Изменения в тематической структуре Живого Журнала: влияние выборов 2011 г.

К. А. Маслинский, О.Ю.Кольцова, С.Н.Кольцов

Статья подготовлена для перевода и сдачи в серию препринтов НИУ-ВШЭ

1 Введение

В последние годы исследователи общественных движений и процессов социальной мобилизации все чаще обращают внимание на роль интернет-медиа в процессах становления и разворачивания социальных движений, прежде всего речь идет о социальных сетях и блогах [Garrett 2006]. В России наиболее яркий эпизод общественной мобилизации, применительно к которому оправданно ставить вопрос о роли интернет-медиа — общественная протестная активность, связанная с проведением и результатами выборов в Государственную Думу в декабре 2011 г. Толчком к данному исследованию послужил вопрос о степени вовлеченности русскоязычной блогосферы в эти процессы: ограничивается ли она отражением выборов и протестных акций наравне с прочими информационными поводами или же блогосфера оказывается интернет-медиа, так или иначе вовлеченном в механизмы социальной мобилизации.

В качестве репрезентативной площадки, представляющей общественные обсуждения в русскоязычном интернете были выбраны наиболее популярные и наиболее читаемые блогеры Живого Журнала. Живой Журнал (LiveJournal.com) выбран в качестве единственной рассматриваемой платформы блог-хостинга, так как это не только исторически основной сервис, на котором возникли русскоязычные блоги [Gorny 2004], но впоследствии также важная публичная площадка для различного рода гражданского активизма [Lonkila 2008]. И хотя с переходом русскоязычной части LiveJournal под контроль российской организации СУП в 2006 г. некоторые авторы предсказывали упадок Живого Журнала как значимой площадки общественно-политических обсуждений в рунете [Parkhomenko, Tait 2008], недавнее масштабное количественное исследование российской блогосферы подтвердило, что большая часть политических блогов, тем не менее, находится там [Этлинг и др. 2010]. В рамках Живого Журнала для исследования выбраны топ-пользователи (по рейтингу Живого Журнала), так как совокупность топ-блогов можно рассматривать как самостоятельное явление, характеризующееся специфической видимостью в в общественном пространстве интернета — повышенной доступностью в топе поиска по блогам и размером аудитории читателей блога (блогеры-«тысячники»).

Идея данного исследования заключается в том, чтобы на основании достаточно репрезентативных количественных данных оценить изменения в тематическом спектре постов Живого Журнала в период общественной активности, последовавшей за выборами

в Государственную Думу в декабре 2011 г. В качестве базы для сравнения может служить «спокойный» период, на несколько месяцев предшествующий изучаемым событиям. В данной работе нас будут интересовать два основных измерения в тематическом спектре блогосферы:

1. *Степень вовлеченности*: насколько много вопрос выборов и связанная с ними протестная активность обсуждается в блогосфере (доля от всех обсуждений в блогосфере); какая доля блоггеров затрагивает эту тему?

2. *Структурные изменения*: изменяют ли общественно-политические темы, связанные с выборами, общую тематическую структуру блогосферы. Например, отражаются ли они на распределении обсуждений, тематически не связанных с выборами (другими словами, вытесняют ли выборы из пространства обсуждений другие, менее значимые темы).

Для того чтобы выводы о тематике обсуждений в Живом Журнале в целом были обоснованными, требуется сбор и тематическая классификация значительных массивов текстов — постов. Масштабы количественных исследований в этой области при использовании классических методов контент-анализа значительно ограничены высокой стоимостью кодирования. В данном исследовании используется полностью автоматический метод анализа содержания текстов с помощью вероятностного тематического моделирования, что позволяет многократно увеличить объемы выборки и скорость получения результатов. Оплата за это — необходимость в достаточно грубых формализациях содержательных структур текста, что может рассматриваться как снижение точности по сравнению с контент-анализом. Тем не менее, если при интерпретации результатов учитывать природу использованных формализаций, данные, полученные методом автоматического анализа, позволяют делать пусть более узкие, но не менее обоснованные утверждения о более широкой выборке текстов.

Дальнейшее изложение в статье будет построено следующим образом. В разделе 2 приводится краткое описание используемого в данной работе алгоритма вероятностного тематического моделирования — Latent Dirichlet Allocation. В разделе 3 описаны принципы построения выборки, процедуры сбора и автоматической обработки текстов, описательная статистика анализируемых данных. В разделе 4 приведены полученные результаты: характеристики выделенных тем и статистика изменений в тематической структуре ЖЖ в декабре 2011 г. Завершает статью раздел 5, в котором подводятся итоги исследования и обсуждаются перспективы дальнейшей работы.

2 Методология

Задача *тематического моделирования* (topic modeling) относится к кругу технологий машинного обучения, применяемых при обработке естественного языка. Цель тематического моделирования — построить статистическую модель, которая описывает

данную коллекцию документов как смесь конечного числа «тем», таким образом что в каждом документе коллекции может быть обнаружено несколько тем, представленных в разных пропорциях. Тема в данном случае взята в кавычки, так как представляет собой ту или иную формальную структуру (вероятностное распределение, формально-языковую модель и т.п.), построенную на основании вычисления частотности и совместной встречаемости слов в текстах коллекции. Качество и полезность того или иного метода тематического моделирования определяется тем, насколько хорошо выделенные «темы» соотносятся с содержательными характеристиками текстов с точки зрения человека.

Модель латентной Дирихле-аллокации (Latent Dirichlet Allocation, далее LDA) была предложена в 2002 г. как метод выделения групп семантически связанных слов в текстовой коллекции [Blei et al. 2003]. В настоящее время LDA и ее модификации являются наиболее широко применяемым методом тематического моделирования для больших текстовых коллекций самых разных жанров: от баз научных статей и художественных текстов до блогов и микроблогов [Blei 2012 ,Blei, Lafferty 2009 ,Ramage et al. 2010].

Модель LDA основана на нескольких формальных упрощениях: каждый текст рассматривается как неупорядоченный набор («мешок») слов; каждое встретившееся в тексте слово считается принадлежащим одной (и только одной) из латентных тем. При этом разные употребления одного и того же слова (например, *митинг*) могут относиться к различным темам. Смысл работы алгоритма в том, чтобы, начав со случайного распределения слов по темам и тем по документам, подобрать такое распределение слов по темам, при котором была бы максимальной вероятность сгенерировать фактически имеющиеся в коллекции тексты, случайно выбирая слова из получившихся тем. На рис. 1 представлен пример тематического разбиения текста, полученного с помощью LDA.

Какие новости ¿1 *? Что в стране происходит* ¿98 *? Что делать* ¿90 *? Кто виноват* ¿90 *? Кто пидарас* ¿90 *?*

Тема	Вес	Топ-слова
Тема 1	0,2	полиция полицейский сотрудник задерживать сообщать милиция находиться **происходить** несколько рассказывать молодой наркотик около
Тема 90	0,6	понимать надо потому **делать** т вообще нужно тоже что-то думать почему про какой-то сейчас
Тема 98	0,2	эфир программа канал журналист немцов передача вести телеканал разговор **новость** интервью тв LJembed радио телевидение москва

Рис. 1**:** Пример тематического анализа текста блога при разбиении на 100 тем. Верхние индексы обозначают принадлежность слова к соответствующей теме

Таким образом, LDA позволяет автоматически разбивать лексику исследуемой коллекции на семантически связанные группы — темы. Слова в рамках темы можно упорядочить по их весу (частотности) в данной теме. Обычно для определения общего содержания темы достаточно просмотреть несколько топ-слов. В реальных текстовых коллекциях темы неравны: могут содержать большее или меньшее число слов, общей или специфичной лексики, в них могут входить слова с большей или меньшей частотностью. Суммарная доля всех словоупотреблений, отнесенных к данной теме, позволяет оценить относительный вес данной темы в рамках коллекции. Также в результате работы LDA для каждого текста вычисляется вектор, характеризующий, в каких пропорциях в данном тексте представлены те или иные темы. Благодаря преобразованию представления текста из вектора слов в вектор тем, LDA можно рассматривать как аналог факторного анализа для текстовых данных [Buntine, Jakulin 2006].

LDA относится к группе методов обучения без учителя, т.е. вычисляет темы полностью автоматически, без предварительной разметки текстов человеком. Основной параметр, который должен быть задан вручную — количество тем. Критерии выбора оптимального числа тем во многом зависят от задач исследования и интерпретируемости результатов: более крупные разбиения позволяют выделить более общие темы, слишком дробное разбиение может отражать случайные черты коллекции. В нашей работе после ряда эмпирических тестов мы остановились на разбиении корпуса постов Живого Журнала на 100 тем.

Анализ эволюции тематики в диахронических текстовых коллекциях (например, в статьях журнала Science за 100 лет) — одна из классических областей применения LDA [Blei, Lafferty 2006]. В данной работе мы не будем включать в модель дополнительных допущений о характере изменения темы во времени, и воспользуемся простым способом оценки диахронических изменений по результатам работы стандартного LDA [Hall et al. 2008]. Для этого сначала выполняется вычисление распределения тем для всей коллекции в целом, затем тексты коллекции разбиваются на временн´ые срезы. В нашей работе каждый срез представляет собой выборку постов, опубликованных в определенном временном интервале (один месяц). Для каждого среза в отдельности вычисляются относительный вес и лексический состав каждой из тем. Сопоставление временных срезов позволяет проследить рост и падение популярности темы (изменения относительного веса темы) и изменения в лексическом составе темы.

Нужно отметить, что при таком подходе нет понятия возникновения или исчезновения тем: все темы являются «сквозными», т.е. представленными во всех временных срезах. Эта особенность модели приводит к тому, что если в коллекции фактически возникают тексты на новую тему, новая лексика этих текстов будет сгруппирована со сходной лексикой текстов других периодов.

Анализ диахронических изменений в тематике блогосферы с помощью LDA подразумевает принятие названных выше формальных упрощений: под изменениями в тематической структуре блогосферы понимаются изменения в суммарной частотности и

распределении по текстам автоматически выделенных групп лексики.

3 Данные

В данном исследовании используется две выборки постов, опубликованных в Живом Журнале с 15 августа по 15 сентября 2011 г. («спокойный период») и с 27 ноября по 27 декабря 2011 г. («активный период»)[1]. Выборку составили посты первых 1400 журналов по рейтингу, опубликованному LiveJournal, на момент сбора данных (октябрь 2011 и январь 2012)[2]. Рейтинг оказался достаточно стабильным: из журналов, вошедших в сентябрьскую выборку 1150 входят и в декабрьскую. В выборку включались посты, опубликованные в журнале в границах заданного периода, но не более 50 первых постов. Для тематического моделирования использованы только тексты постов (без комментариев). Предварительная обработка текстов включала лемматизацию с помощью программы mystem [Segalovich] с автоматическим выбором более частотной либо первой леммы, удаление 100 наиболее частотных слов коллекции и слов, встречающихся менее чем в 5 документах, а также документов, содержавших менее 5 слов после обработки. После обработки в сентябрьской выборке 24,198 постов 1360 пользователей; в декабрьской — 27,026 постов 1393 пользователей. LDA-разбиение на 100 тем и последующий анализ веса и лексического состава тем в каждом из периодов выполнялись с помощью Stanford Topic Modeling Toolbox [Ramage et al. 2009].

4 Результаты

4.1 Характеристика выделенных тем

Автоматически выделенные LDA темы были закодированы вручную на основании просмотра топ-20 слов темы и топ-20 текстов, отнесенных к данной теме с наибольшей вероятностью. При кодировании каждой теме приписывалась метка (слово, характеризующее тему), название и тип. В данной работе использована эмпирическая классификация тем на 4 класса по характеру топ-лексики и топ-текстов:

3. Предметные — темы с выраженным ядром лексики, относящейся к определенной предметной области, примеры — *здоровье* (медицина и здоровье), *СССР* (советская история и реалии).

4. Стилистические — темы с преобладанием в топе общей лексики, характеризующие в большей степени стилистику (регистр) текста и не ассоциированные с узкой

---

1 Данные собраны с помощью программы Koltran Blogminer, разработанной в Лаборатории Интернет-Исследований Высшей Школы Экономики, Санкт-Петербург.

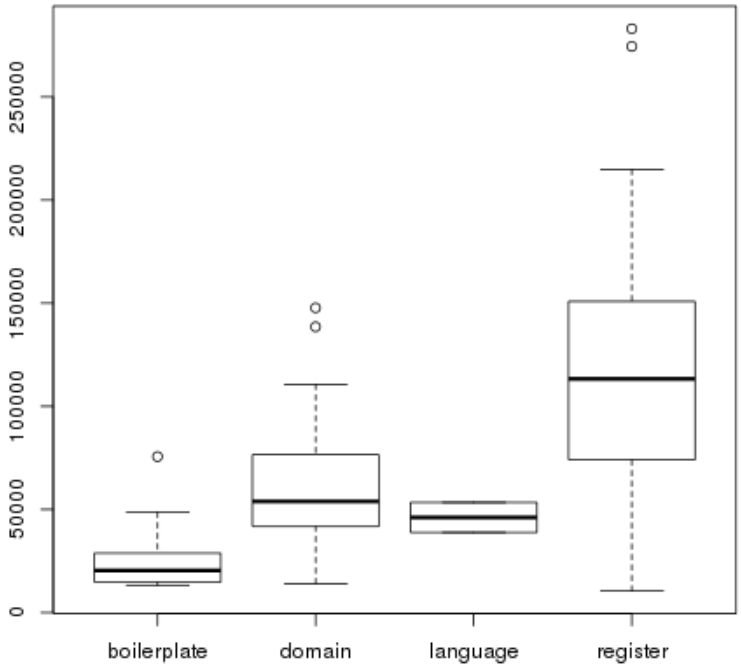
2 Рейтинг LiveJournal: URL. До февраля 2012 г. рейтинг составлялся только на основании общего количества пользователей, включивших данный журнал в список «друзей».

предметной областью; примеры — *сегодня* (сиюминутные записи, планы на будущее), *надо* (суждения, модальность).

5. Языковые — темы, ядро лексики которых составляют слова на иностранном языке (в нашей коллекции это английский и украинский).

6. Формальные — темы, сформировавшиеся на основе клишированных фраз, зачастую текстовых элементов интерфейса, списочных структур и т.п., а также все темы, не поддающиеся интерпретации (статистический шум).

В результате из 100 тем 65 были отнесены к предметным, 21 — к стилистическим, 2 — к языковым и 12 — к формальным. Нужно отметить, что границы между названными категориями очень размыты, поэтому отнесение тем к тому или иному классу достаточно условно. Тем не менее, различные типы тем в целом различаются по таким характеристикам как общее число слов в теме и изменчивость лексики в топе темы при сравнении сентябрьского и декабрьского срезов. Стилистические темы в среднем содержат больше слов и их лексическое ядро более стабильно в диахронии по сравнению с предметными темами, что хорошо объясняется тем, что они преимущественно состоят из общей лексики (см. рис. 2).



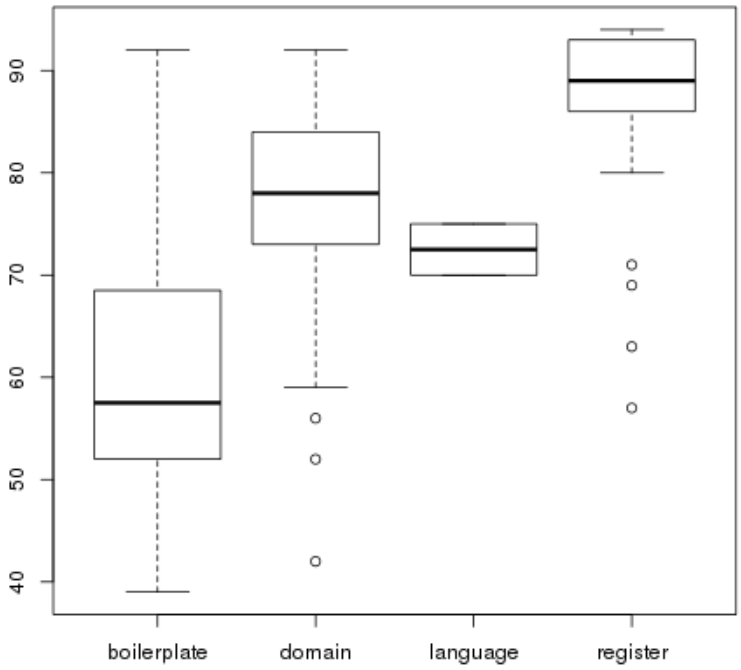


Рис. 2: Количественные характеристики разных типов тем

Для предметных тем характерна стабильность на уровне около 80 из 100 топ-слов в декабрьской выборке по сравнению с сентябрьской. Например, в теме *церковь* 79 общих слов в списке топ-100 для сентября и декабря, в частности *церковь, храм, православный, бог, святой, Христос* и т.п. Среди слов, специфических для топ-100 в сентябре — *Охлобыстин, Чаплин, иудей, Иерусалим* и др., в декабре — *католик, рождество, спаситель, святыня* и др. Этот пример показывает, что выделенные темы не позволяют судить о составе «повестки дня», обсуждаемой в блогосфере, так как упоминания конкретных событий и персонажей встраиваются в тему, основу которой скорее составляет более частотная и более общая лексика определенной предметной области, такой как православная церковь, военные действия (тема *Ливия*) или судебные разбирательства (тема *суд*). Представляется, что корректным уровнем интерпретации результатов LDA по текстам блогосферы будет оценка количественных трендов в частотности упоминания тех или иных предметных областей.

4.2 Динамика популярности тем

Самый простой способ оценки изменений в тематическом составе блогосферы — сравнение относительного веса тем, т.е. расчет доли текстовой коллекции, посвященной данной теме в каждый из периодов. Поскольку в результате работы LDA каждый текст

разделяется между несколькими темами в разных пропорциях, естественно говорить не о количестве постов на определенную тему, а о доле данной темы в каждом из постов. Таким образом, относительный вес темы можно вычислить, просуммировав доли темы в каждом тексте и поделив на общее количество текстов. Полученное число можно рассматривать как общую долю, занимаемую темой в текстовой коллекции, нормализованную по количеству постов и по длине поста, т.к. один и тот же вес темы в длинном и коротком тексте будет фактически соответствовать разному количеству слов. Например, в посте из 5 слов тема с весом 0,2 соответствует одному слову, а в посте из 500 слов тот же вес соответствует 100 словам, отнесенным к данной теме. Нормализация по длине текста позволяет учесть тот факт, что пост является коммуникативной единицей блогосферы, и позволяет вычислить тенденцию высказываться на ту или иную тему, не учитывая фактор большей или меньшей многословности авторов.

Другой способ расчета относительного веса темы предполагает нормализацию только по количеству слов в коллекции, в этом случае подсчитывается суммарное количество слов, отнесенных к данной теме во всех текстах. При этом темы, обсуждаемые в более пространных текстах, получат больший относительный вес по сравнению с темами, характерными для более коротких текстов.

Изменения в относительном весе тем в декабрьской коллекции по сравнению с сентябрьской, рассчитанные в долях текстов и в долях слов, показывают высокую корреляцию друг с другом ($r = 0{,}94$). Очень резких изменений в объеме тем в наших данных не засвидетельствовано: в долях текстов разброс значений от $-0{,}65$ до $1{,}92$ (медиана $-0{,}08$), в долях слов от $-0{,}87$ до $1{,}86$ (медиана $-0{,}06$). При этом отчетливо наблюдается тенденция к общему относительно равномерному падению веса большей части тем на фоне более выраженного роста нескольких тем (рис. 3). Такая картина может интерпретироваться как относительное снижение тематического разнообразия в декабре.

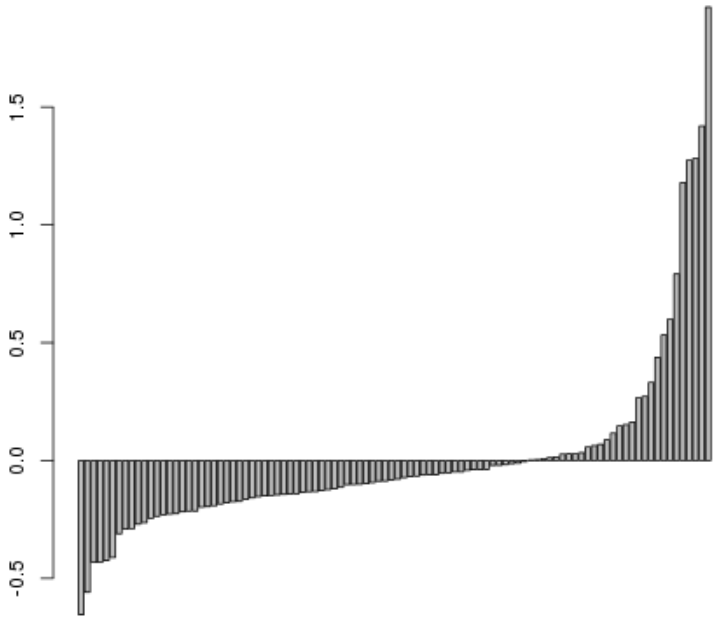


Рис. 3: Изменения веса тем в декабре по сравнению с сентябрем (в процентах от количества документов в коллекции)

Список тем, относительный вес которых в наибольшей степени сократился в декабре, приведен в таблице 1. Часть изменений может быть объяснена тем, что ядро темы составляет лексика, в рамках данной коллекции тесно ассоциированная с определенными событиями (новостными поводами), потерявшими актуальность к декабрю. Таковы темы *Ливия* (военные события в Ливии и других арабских странах) и *самолет* (авиационная лексика, прежде всего в связи с крушением самолета Як-42 в Ярославле). Изменения в других темах могут быть объяснены сезонным спадом актуальности, таковы *туризм* и *школа* (1 сентября). Причиной прочих изменений могут быть как случайные флуктуации, так и более сложные содержательные процессы, не имеющие простой интерпретации на уровне семантических групп лексики.

Таблица 1: Темы, сократившиеся в декабре

№	метканазвание	в документах	в словах	тип
52	Ливияреволюционные события на Ближнем Востоке	-0.653430777	-0.868439973	domain
40	туризмотчеты о путешествиях	-0.558579537	-0.534375172	domain
50	тюрьмакриминальная	-0.431869292	-0.620776838	domain

	хроника			
89	сегоднясиюминутные записи	-0.431260913	-0.280361042	register
92	самолетавиация, авиакатастрофы	-0.423653326	-0.331311712	domain
39	улицагородская среда и архитектура	-0.411627346	-0.506240318	domain
84	футболновости о футболе	-0.310366018	-0.282785753	domain
82	школашкольное образование	-0.290695286	-0.283097334	domain
94	устройствокомпьютеры и мобильные устройства	-0.289248778	-0.270207407	domain
76	семьясемья и дети	-0.270033014	-0.250015726	domain
19	календарьназвания дней недели, списки событий	-0.263461360	-0.039178651	boilerplate
21	поэзиястихотворные тексты	-0.244954115	-0.345416911	register
73	Европаназвания европейских стран	-0.236339630	-0.152887207	domain
56	концертмузыка и концерты	-0.229841629	-0.229359544	domain
6	войнавоенные действия	-0.227134570	-0.449049135	domain

Список тем, относительный вес которых вырос в декабре в наибольшей степени приведен в таблице 2. Обращает на себя внимание, что в вершине этого списка доминируют темы, непосредственно ассоциированные с послевыборными протестными акциями (*митинг, площадь, революция*), политическими партиями и результатами голосования (*голос, партия, участок*), законодательством (*закон*) и государственной властью вообще (*власть, путин*). Единственным исключением является тема *подарок*, связанная с обсуждением новогодних подарков (конец рассматриваемого периода — 27 декабря), причем относительно более низкий прирост этой темы в словах показывает, что она встречается в более коротких текстах. Напротив, тема *участок*, связанная с обсуждением процесса и результатов голосования на отдельных избирательных участках, демонстрирует более выраженный прирост в словах, т.е. тенденцию встречаться в более пространных текстах.

Примечательно, что среди подавляющего большинства предметных тем, заметный рост демонстрируют как минимум две стилистические темы. Первая из них — *надо* (суждения, модальность), ее рост в вероятнее всего определенную стилистику и коммуникативные функции основной массы текстов на значительно расширившиеся политические темы. Рост темы *поделиться* может свидетельствовать об активизации перепостов, т.е. актуализации транслирующей информационной функции блогов.

№	метканазвание	в документах	в словах	тип
22	митингпротестные акции в Москве	1.923797381	1.762881477	domain

27	площадьпротестные акции и аресты	1.419119175	1.158872722	domain
68	революцияпротестные акции	1.281907478	1.578458132	domain
55	подарокподарки к новому году и другим праздникам	1.274445226	0.770962171	domain
33	голосрезультаты голосования	1.179473817	1.449827305	domain
3	партияполитические партии	0.791644505	0.911135968	domain
72	надосуждения, модальность	0.599963357	0.707968929	register
66	властьобщеполитическая лексика, государство	0.532548434	0.821218859	domain
90	путинупоминания Путина и Медведева	0.437745111	0.345851988	domain
65	поделитьсяперепосты: «поделиться» в социальных сетях	0.331676898	0.178751695	register
57	Ким Чен Ирсмерть Ким Чен Ира	0.272162870	0.149954110	domain
83	законзаконодательство, ЦИК	0.266384868	0.414105169	domain
60	каналсми и интервью	0.163111704	0.018782701	domain
34	postedперепосты: «posted via»	0.152127501	0.141651184	boilerplate
69	рейтингперепосты, блоги	0.146538799	-0.003749349	register
25	кинокино и театр	0.115927864	0.025382394	domain
38	участокподсчет голосов на выборах, наблюдатели	0.087651463	0.463976424	domain

Таблица 2: Темы, расширившиеся в декабре

Показателем стабильности содержания Живого Журнала вне зависимости от сезонных и

политических обстоятельств могут служить темы, относительный вес которых изменился в декабре в наименьшей степени (таблица 3). Вершину этого списка составляют темы, связанные с повседневностью пользователя интернета (*деньги* — зарплаты, платежи; *связь* — мобильная связь и интернет), а также с его хобби (*фотография*). Также в этом списке значительно более широко, чем в предыдущих, представлены формальные и стилистические темы, т.к. они сформированы в основном общей лексикой или формальными структурами в тексте, для которых нет оснований ожидать значительных частотных колебаний.

метка	название	в документах	в словах	тип
деньги	зарплаты, платежи, денежные суммы	0.002469855	-0.067491861	domain
связь	мобильная связь и интернет	0.003868197	0.130860743	domain
фотография	фотография и фотографы	-0.003877876	0.106742732	domain
false	шум	0.007842570	0.129713763	boilerplate
Украина	украинские новости	-0.010816311	-0.066030108	domain
газета	журналистика: газеты, статьи	0.012415853	0.086415612	domain
суд	судебные процессы	-0.012757500	0.002717530	domain
кнопка	названия элементов интерфейса	0.015136251	0.104444706	boilerplate
изображение	ссылки на изображения	-0.017578742	-0.033626703	boilerplate
случай	описания случаев, нарративы	-0.019676473	0.093355865	register
ресторан	продукты и напитки: кофе, вино, пиво, водка	-0.020121581	0.018173527	domain
процент	раскладки в процентах	0.026916010	-0.178197499	register
видеть	общая лексика: глаголы и союзы	0.028338709	0.060572558	register
ролик	досуговые записи: ссылки на интересные материалы	0.028494774	-0.016400204	register
дом	квартира и домашнее хозяйство	0.033226168	-0.003902871	domain
конкурс	публичные мероприятия: конкурсы	-0.036977114	0.066848107	domain
швабра	шум: разные тексты и лексика	-0.038167524	-0.016251757	boilerplate
глава	новости о чиновниках	-0.038674127	-0.022334798	domain
smile	общая разговорная лексика	-0.041145085	-0.091634214	register
животные	названия животных	-0.048040134	-0.081535053	domain

Таблица 3: Стабильные темы

4.3 Распределение тем по авторам

Другое возможное измерение тематической структуры блогосферы предполагает анализ характера распределения тех или иных тем по авторам. Заметные изменения в таком

распределении можно рассматривать как структурное изменение в содержательном поле блогосферы. В данном разделе анализируется посты только тех 1122 авторов, которые присутствуют в выборке по декабрьскому и сентябрьскому рейтингу и имеют в каждом из периодов не менее одного поста.

Так как основной интересующий нас вопрос — возможные изменения в структуре обсуждения политических событий в Живом Журнале, из 100 выделенных LDA тем мы отобрали 11 тем, так или иначе ассоциированных с обсуждением внутриполитических событий и действий государственных властей: *суд*, *партия*, *митинг*, *площадь*, *участок*, *тюрьма*, *власть*, *революция*, *закон*, *администрация*, *Путин*. Для определения постов, в значительной степени связанных с внутриполитическими событиями мы использовали установленный эмпирически пороговый критерий — при суммарной доле «внутриполитических» тем в тексте более 0,3 текст относился к разряду «политических», в противном случае он классифицировался как «неполитический». Нужно заметить, что этот критерий нельзя рассматривать как способ оценки заинтересованности автора поста во внутриполитических событиях, что связано не только с особенностями работы LDA, но и весьма условным понятием «внутриполитического», в которое мы не включили темы, связанные с более широкими общественными вопросами, такие как *образование* или *общество*. Однако этот подход можно рассматривать как способ грубой оценки количества журналов и постов, в которых присутствуют неслучайные употребления слов, ассоциированных с внутриполитическими событиями.

Статистика по общему количеству постов приведена в таблице 4. На фоне общего роста среднего количества постов также несколько выросло и среднее количество «политических» и несколько упало среднее количество «неполитических» постов одного уровня. При интерпретации роста количества «политических» постов необходимо учитывать, что значительная часть выделенных нами «внутриполитических» тем тесно ассоциированы с новостными поводами, возникшими только в декабре. Поэтому в сентябрьской части коллекции следует ожидать только текстов, отнесенных к этим темам преимущественно на основании периферийной лексики либо случайных совпадений.

Постов автора	сентябрь		декабрь	
	среднее	медиана	среднее	медиана
Всего	19,42	14	20,14	15
«Политических»	1,62	0	3,74	1
«Неполитических»	17,80	13	16,46	12

Таблица 4**:** Статистика по количеству политических и неполитических постов

С другой стороны, в наших данных нет определенных свидетельств, что прирост общего количества постов в декабре объясняется приростом постов на политические темы. В целом блогеры демонстрируют устойчивое поведение в отношении частотности написания постов: корреляция между количеством постов блогера в сентябре и декабре довольно высока (0,80). Наблюдается умеренная положительная корреляция (0,40) между изменением общего количества постов и изменением количества политических постов

данного автора. Однако значительно более выражена корреляция между изменением общего количества постов и изменением количества «неполитических» постов (0,88). Вполне вероятно, что общий рост количества постов в декабре объясняется в первую очередь сезонным ростом активности блогеров.

Распределение политических постов по всей выборке пользователей также демонстрирует определенную динамику в декабре по сравнению с сентябрем. Так, в сентябре 683 пользователя (60,9%) не имеют «политических» постов, в то время как в декабре — только 511 (45,5%). При этом у 228 (20%) пользователей, не имевших «политических» постов в сентябре, они появляются в декабре, напротив, только у 56 (5%) пользователей политические посты в декабре исчезают. В целом эти данные показывают вовлечение большего числа блогеров в обсуждение внутриполитических вопросов.

5 Заключение

Нам представляется, что результаты, полученные с помощью автоматического лексического анализа по методике LDA, позволяют говорить о некоторых структурных изменениях в тематике топа Живого Журнала в декабре 2011 г. Прежде всего наблюдается существенный рост веса всех тем, тесно ассоциированных с обсуждением результатов выборов и протестной активности, на фоне более умеренного спада, затрагивающего подавляющее большинство оставшихся тем. Таким образом, мы наблюдаем «фокализацию» тематики блогов на проблемах, связанных с выборами и последовавшей гражданской активностью. В наших данных имеются также косвенные свидетельства актуализации ретранслирующей функции блогов в декабре (повышенная частотность перепостов).

В декабре также прослеживается рост числа пользователей, в постах которых значительный вес получает лексика, составляющая «внутриполитические» темы. Приток пользователей, чьи посты можно условно классифицировать как «политические» по данным LDA, значительно превышает отток, что может свидетельствовать о процессах мобилизации блогеров в политической тематике. В то же время гипотеза о том, что общая активность блогосферы (измеряемая количеством постов) возросла за счет прироста постов на политические темы, не подтверждается нашими данными. Однако в силу методологических ограничений предложенного критерия классификации постов, эти данные можно считать лишь очень предварительными и требующими дальнейшей проверки.

В целом, на настоящем этапе исследования недостаточно данных, для того чтобы оценивать масштаб обнаруженных структурных изменений в тематике Живого Журнала. Необходимы дальнейшие диахронические исследования, которые позволили бы установить характерные для Живого Журнала ритмы смена тем и пределы нормальных колебаний в широте набора актуальных тем. Подобный сравнительный материал позволит отличать сезонные колебания активности в Живом Журнале от существенных тематических изменений, спровоцированных внешними социальными процессами, чтобы

приблизиться к ответу на вопрос, какую роль в этих процессах может играть русская блогосфера в целом и Живой Журнал в частности.

Список литературы

[Blei 2012] *D. Blei*. Probabilistic topic models. Communications of the ACM, 55(4):77–84, 2012.

[Blei, Lafferty 2006] *D. Blei and J. Lafferty*. Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning, 2006.

[Blei, Lafferty 2009] *D. Blei and J. Lafferty*. Topic Models. In A. Srivastava and M. Sahami, editors, Text Mining: Classification, Clustering, and Applications . Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2009.

[Blei et al. 2003] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John. Latent Dirichlet allo- cation. Journal of Machine Learning Research 3: pp. 993–1022. doi:10.1162, 2003.

[Buntine, Jakulin 2006] *W. Buntine and A. Jakulin*. Discrete component analysis. In Subspace, Latent Structure and Feature Selection. Springer, 2006.

[Garrett 2006] *Garrett, R. Kelly* Protest in an Information Society: a review of literature on social movements and new ICTs // Information, Communication & Society, 2006. Vol. 9. № 02. Pp. 202—224.

[Gorny 2004] *Gorny E.* Russian LiveJournal: National specifics in the Development of a Virtual Community. Version 1.0 of 13 May 2004. Russian-cyberspace.org http://www.ruhr-uni-bochum.de/russ-cyb/library/texts/en/gorny_rlj.pdf. (вход выполнен 05.04.2012).

[Hall et al. 2008] *Hall, D. and Jurafsky, D. and Manning, C.D.*, Studying the history of ideas using topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. 363–371. 2008.

[Lonkila 2008] *Lonkila, Markku* (2008): The Internet and Anti-military Activism in Russia, Europe-Asia Studies, 60:7, 1125—1149.

[Parkhomenko, Tait 2008] *Parkhomenko Y., Tait A.* Blog Talk // Index on Censorship. February 2008 37: 174-178, doi:10.1080/03064220701882822. http://ioc.sagepub.com/content/37/1/174.citation (Вход выполнен 08.06.2012).

[Ramage et al. 2009] *Ramage D., Rosen E., Chuang J., Manning C.D., McFarland D.A.* Topic Modeling for the Social Sciences. NIPS 2009 Workshop on Applications for Topic Models (Вход выполнен 19.04.2012).

[Ramage et al. 2010] *Ramage D., Dumais S., Liebling D.* Characterising Microblogs with Topic Models. ICWSM 2010. http://www.stanford.edu/ dramage/papers/twitter-icwsm10.pdf (Вход выполнен 19.04.2012).

[Segalovich] *Segalovich, I.* A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. // Proceedings of MLMTA. 2003. http://download.yandex.ru/company/iseg-las-vegas.pdf (Вход выполнен 08.06.2012).

[Этлинг и др. 2010] *Этлинг Б., Алексанян К., Келли Дж., Палфри Дж., Гассер У.* Публичный дискурс в российской блогосфере: анализ публичной политики и мобилизации. Исследования центра Беркмана No 2010-11, 19 октября 2010. http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Public_Discourse_in_the_Russian_Blogosphere-RUSSIAN.pdf. English language original: http://cyber.law.harvard.edu/publications/2010/Public_Discourse_Russian_Blogosphere. (Вход выполнен 17.04.2012).