

Sentiment analysis комментариев к постам в ЖЖ

Руководитель: О.Кольцова

Участники: К.Маслинский, Ю.Павлова

Sentiment analysis

- Sentiment analysis – одно из направлений текстового анализа, главной целью которого является определение эмоциональной окраски текста.
- При определении оценок чаще всего используется поляризованная шкала: позитивная—негативная оценка
- Основными источниками информации для автоматического определения оценок в тексте является прежде всего лексика, также может учитываться пунктуация и специальные конвенции.

Эмоции в блогах

- Изменения в эмоциональной окраске комментариев и постов связано с каким-либо событием, темой, социально значимой проблемой.
- Эмоции автора способны влиять на читателей => могут изменить их представления о явлении или подтолкнуть к реальным действиям.

Привееет!! =) Вчера
отлично провели время
в кино =) Фильм очень
понравился=)

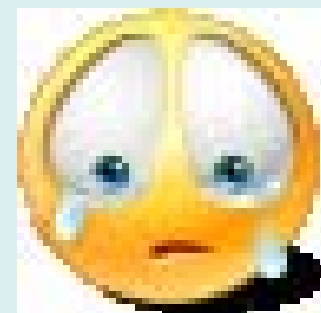
Нарушения на прошедших
выборах в гос.думу ! С этим
надо бороться! Это надо
пресекать!

Давайте пойдем и покажем
этим ворам, что мы не рабы!
И уничтожим всех, кто встанет
на нашем пути!

- Гипотеза – если существует какая-то социальная напряженность в обществе по отношению к какой-либо теме, то это находит выражение в постах и комментариях блогов.
- Задача исследования – найти темы в блогах, которые вызывают социальную напряженность.

SentiStrength

- Проф. Майк Фелвол, глава Statistical Cybermetrics Research Group университета Вулверэмптона и ассоциированный научный сотрудник Oxford Internet Institute, Великобритания.
- SentiStrength оценивает силу позитивных и негативных эмоций на основе словаря, все слова в котором закодированы:
 - 1 (not negative) to -5 (extremely negative)
 - 1 (not positive) to 5 (extremely positive)



SentiStrength

Адаптация SentiStrength

- Перевод словаря SS на русский язык
- Подбор подходящих русских эквивалентов
- Составление частотного словаря слов по комментариям ЖЖ
- Объединение этих словарей
- Кодирование словаря

Наша база:

- 2 коллекции постов блогов с комментариями топ-2000 ЖЖ
- Два временных периода: 15 августа-15 сентября, 27 ноября - 27 декабря
- Выборка по двум основам ислам* и мусульман* (со всеми производными формами)
- Случайная выборка, не касающаяся ислама, для проверки пригодности ПО к другим темам

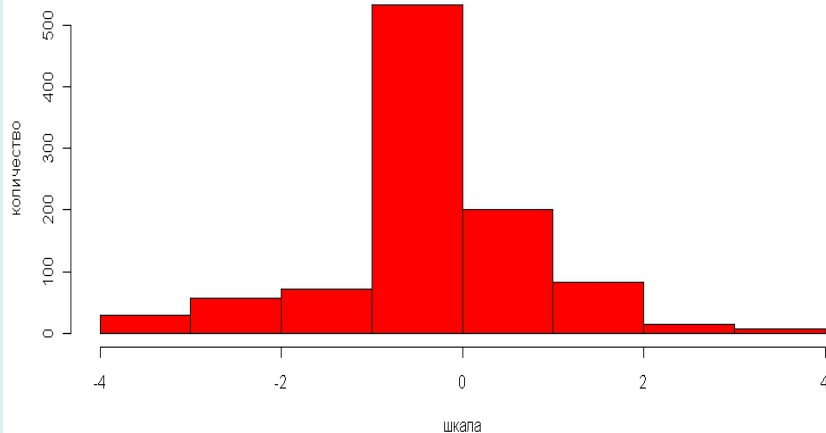
Единица анализа - комментарий

Дизайн исследования

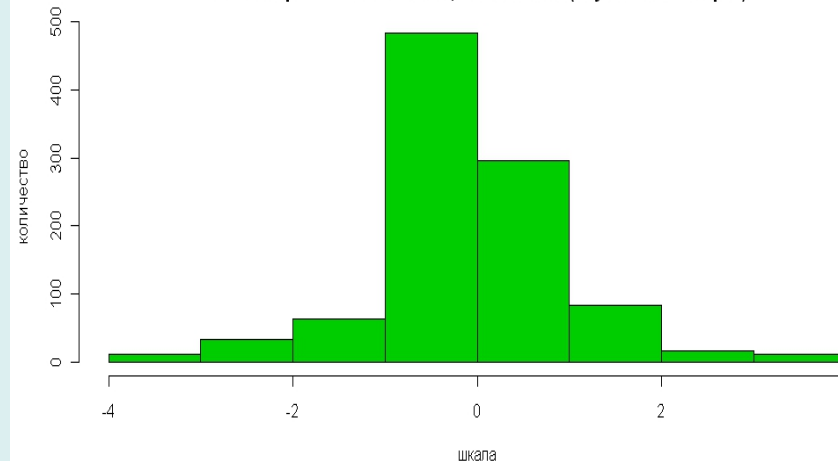
- Оцениваем эмоциональную заряженность комментариев к постам по выборкам об «исламе» и случайной выборке с помощью SentiStrength.
- Оцениваем эмоциональную заряженность комментариев к постам по выборкам об «исламе» и случайной выборке с помощью кодировщиков.
- Сравниваем результат, делаем вывод о работоспособности инструмента.
- Сравниваем результат, делаем вывод о наличии социальной напряженности по отношению к теме ислама.

Автоматический анализ

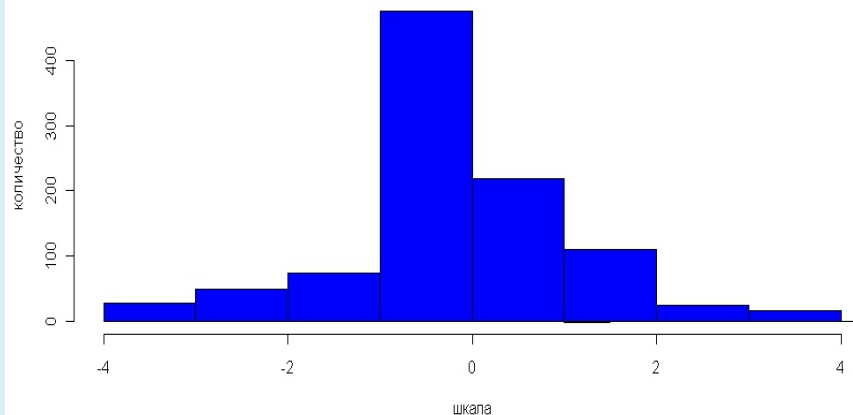
Распределение
комментариев по шкале эмоциональности (сентябрь)



Распределение
комментариев по шкале эмоциональности (случайная выборка)



Распределение
комментариев по шкале эмоциональности (декабрь)



Сравнение с ручным кодированием. Декабрь

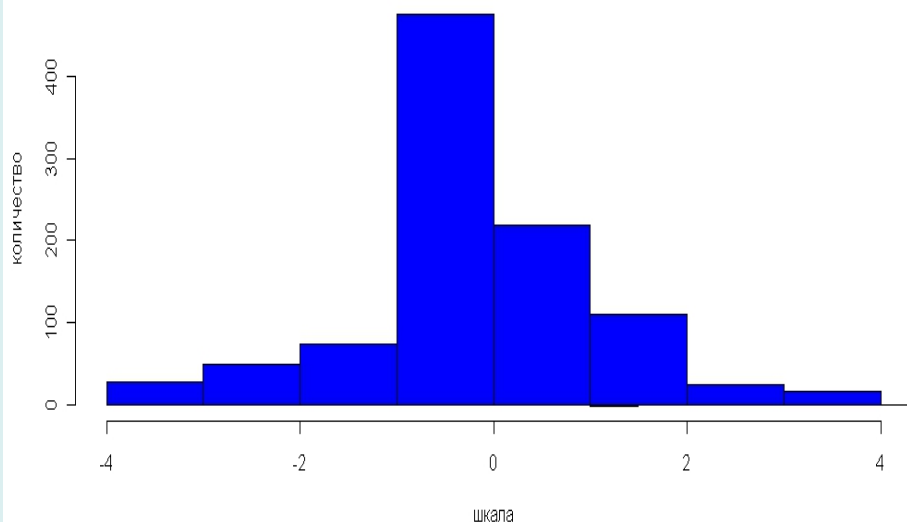
Автоматический анализ

-5	-4	-3	-2	-1	0	1	2	3	4	5
	7	21	50	74	476	219	111	25	17	
0										0

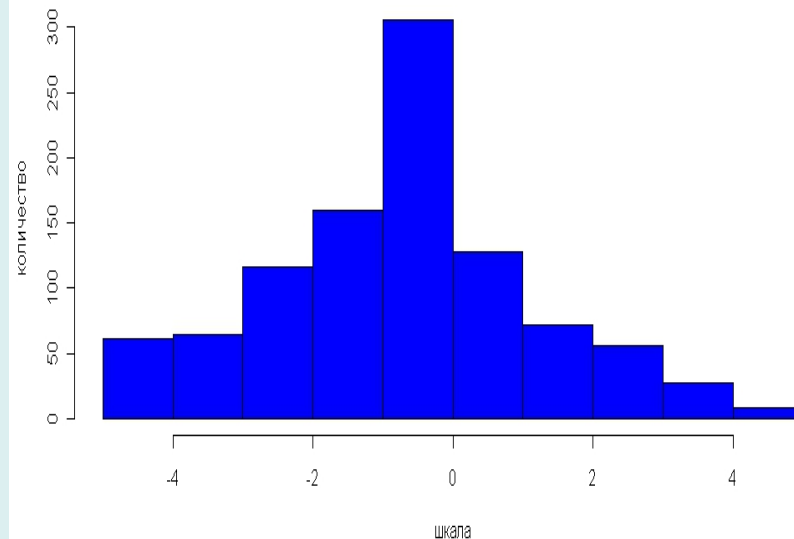
Кодировщики:

-5	-4	-3	-2	-1	0	1	2	3	4	5
	39	65	116	160	306	128	72	56	27	
22										9

Распределение
комментариев по шкале эмоциональности (декабрь)



Распределение
комментариев по шкале эмоциональности Coders (Декабрь)



Сравнение с ручным кодированием.

Случайная выборка

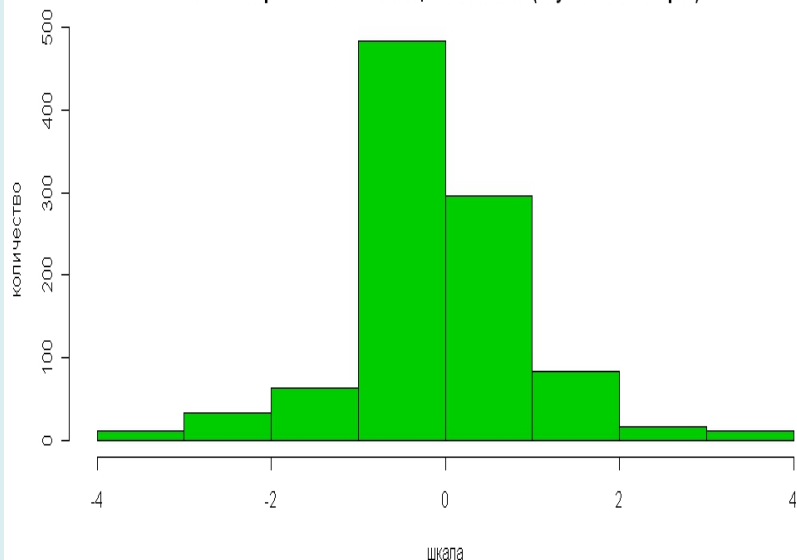
Автоматический анализ:

-5	-4	-3	-2	-1	0	1	2	3	4	5
	4	8	33	63	484	294	84	16	11	
0										0

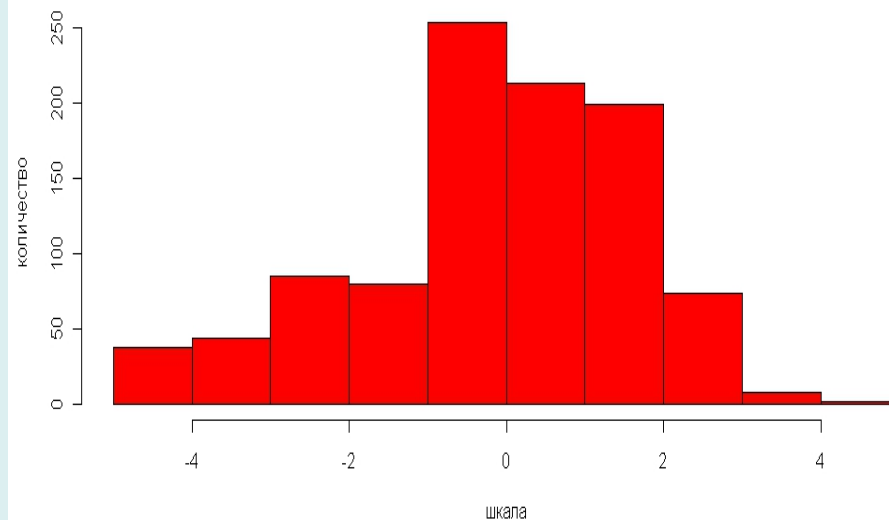
Кодировщики:

-5	-4	-3	-2	-1	0	1	2	3	4	5
	29	44	85	80	254	213	199	74	8	
2										9

Распределение
комментариев по шкале эмоциональности (случайная выборка)



Распределение
комментариев по шкале эмоциональности Coders (Случайная выборка)



Сравнение с ручным кодированием. Сентябрь

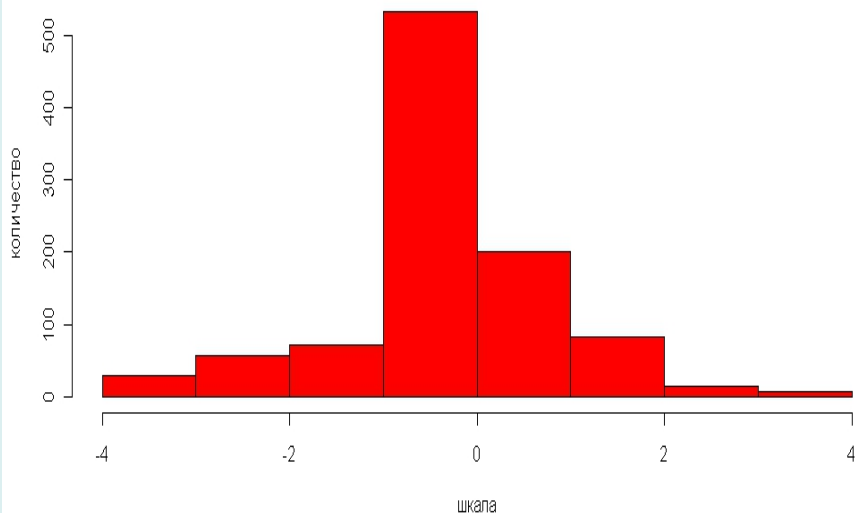
Автоматический анализ

Кодировщики

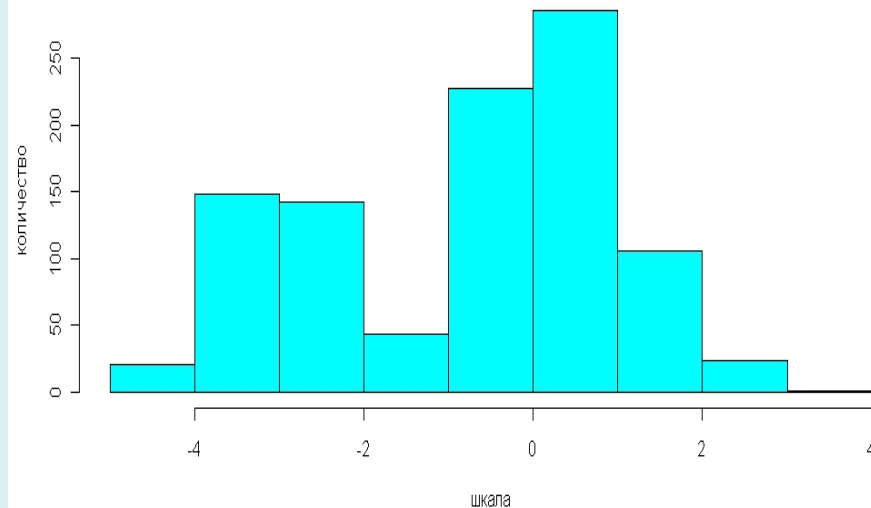
-5	-4	-3	-2	-1	0	1	2	3	4	5
0	7	20	50	74	476	219	111	25	17	0

-5	-4	-3	-2	-1	0	1	2	3	4	5
6	15	148	142	44	227	286	106	24	1	0

Распределение
комментариев по шкале эмоциональности (сентябрь)



Распределение
комментариев по шкале эмоциональности Coders (Сентябрь)



Оценка качества

	% совпадений кол-ва оценок	% точного совпадения оценок	% точного совпадения оценок+ (+/-1)
Сентябрь	60,20%	35,70%	64,60%
Декабрь	60%	29,10%	60,80%
Случайная выборка	62,80%	24%	63,60%

Выводы:

- Социальная напряженность по отношению к теме ислама (сентябрь), выявленная с помощью ручного кодирования.
- Необходимость доработки инструмента для автоматического анализа.
- Улучшение качества словаря.

Спасибо за внимание!