Градуировка тональной шкалы реакций на интернет-новости (на материале русского, английского и испанского языков)

Образовательная программа ВМ.5626
«Прикладная и экспериментальная лингвистика»
Профиль «Компьютерная лингвистика и интеллектуальные технологии»

Савинцева Дарья Викторовна

Научный руководитель: к. ф. н., доц. Азарова И. В.

16 июня 2020 г.

Цель и задачи работы

Цель

Создать тональную шкалу тем, представленных в комментариях пользователей социальных сетей на актуальные новости.

Задачи

1. Определить рабочую модель тонального анализа на базе современных методов автоматического анализа текстов:
   - Подобрать и дополнить существующие тональные лексиконы для русского, английского и испанского языков;
   - Определить модель тематического моделирования;
2. Создать корпусы комментариев в новостных группах социальных сетей на русском, английском и испанском языках;
3. Разработать процедуру выделения тональных компонентов в текстах, используя тематическое моделирование и тональные лексиконы;
4. Разработать шкалу для сопоставления тональной окраски тем.

Анализ тональности и тематическое моделирование

Анализ тональности

- Это набор методов выделения эмоционально-экспрессивно окрашенных компонентов в тексте.
- В исследовании используется словарный подход.

Тематическое моделирование

- Это инструмент статистического анализа текста, задача которого выделить темы в коллекции документов, то, как эти темы связаны друг с другом.
- В исследовании используется алгоритм тематического моделирования латентное размещение Дирихле, реализованный в библиотеке gensim.

Гипотеза

Распределение тональных компонентов в темах комментариев может показать:

- насколько тонально окрашена та или иная тема в сравнении с остальными;
- какова средняя тональная окраска темы;
- разницу в тональной окраске тем в разных языках.

Так станет возможно расположить темы на тональной шкале, основанной на тональности комментариев пользователей.

Корпусы комментариев

Язык	Источники (кол-во)	Комментарии	Словарь	
			до обработки	после обработки
русский	Вконтакте (47)	6,022,501	2,003,523	557,091
английский	YouTube (15)	6,463,610	1,158,766	399,738
испанский	YouTube (15)	526,436	284,714	101,858

Тональный словарь для русского языка

- создан на основе словаря Linis-crowd
- 11,653 единицы
- -1 или 1 тональность

```
подвиг,Nn,1
подвижник,Nn,1
подвижница,Nn,1
подвижнический,Aj,1
подвижничество,Nn,1
подвижность,Nn,1
подвижный,Aj,1
подводить,Vb,-1
подворовывание,Nn,-1
подворовывать,Vb,-1
подвох,Nn,-1
```

Тональный словарь для английского языка

- Opinion (Sentiment) Lexicon
- 6789 единиц
- -1 или 1 тональность

```
outsider,-1
outsmart,1
outstanding,1
outstandingly,1
outstrip,1
outwit,1
ovation,1
over-acted,-1
over-awe,-1
over-balanced,-1
over-hyped,-1
```

Тональный словарь для испанского языка

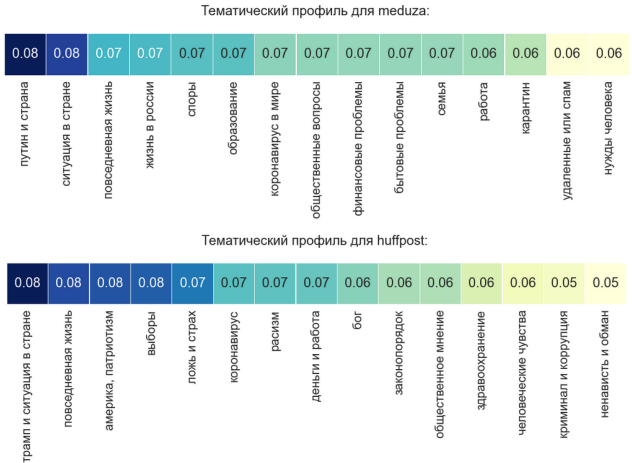
- создан на основе словарей improved Spanish Opinion Lexicon и Spanish Sentiment Lexicon
- 10,463 единицы
- -1 или 1 тональность

```
tirana,-1
tiranas,-1
tirania,-1
tiranica,-1
tiranicamente,-1
tiranicas,-1
tiranico,-1
tiranicos,-1
tiranizar,-1
tirano,-1
tiranos,-1
```

Выделенные темы в корпусах

Русский	Американский	Латиноамериканский
общественные вопросы	общественное мнение	демократия
ситуация в стране	ложь и страх	венесуэла и россия
бытовые проблемы	здравоохранение	революционные настроения
споры	коронавирус	правительство
нужды человека	бог	воровство и терроризм
путин и страна	трамп и ситуация в стране	президент
работа	человеческие чувства	проблемы людей
карантин	расизм	освобождение
финансовые проблемы	законопорядок	ложь и предательство
семья	деньги и работа	ожидания
жизнь в россии	ненависть и обман	ситуация в мире
коронавирус в мире	повседневная жизнь	базовые нужды
удалённые или спам	америка, патриотизм	сша и латинская америка
повседневная жизнь	выборы	коммунисты и диктатура
образование	криминал и коррупция	трамп

Визуализация полученных тем



Распределение тональных компонентов в российском корпусе

Topic	sentim
повседневная жизнь	51
бытовые проблемы	47
жизнь в россии	47
общественные вопросы	45
работа	41
ситуация в стране	38
нужды человека	36
споры	36
семья	35
финансовые проблемы	35
образование	34
карантин	32
путин и страна	31
удаленные или спам	31
коронавирус в мире	29

Положительные и отрицательные

Topic	sentim
бытовые проблемы	40
жизнь в россии	37
повседневная жизнь	36
общественные вопросы	32
работа	32
ситуация в стране	31
нужды человека	25
образование	25
споры	25
семья	23
удаленные или спам	22
финансовые проблемы	22
коронавирус в мире	21
карантин	20
путин и страна	20

Отрицательные

Topic	sentim
повседневная жизнь	15
общественные вопросы	13
финансовые проблемы	13
карантин	12
семья	12
нужды человека	11
путин и страна	11
споры	11
жизнь в россии	10
образование	9
работа	9
удаленные или спам	9
коронавирус в мире	8
бытовые проблемы	7
ситуация в стране	7

Положительные

Распределение тональных компонентов в американском корпусе

Topic	sentim
ложь и страх	45
бог	40
здравоохранение	40
ненависть и обман	38
законопорядок	37
повседневная жизнь	36
коронавирус	34
расизм	33
криминал и коррупция	32
общественное мнение	32
человеческие чувства	31
америка, патриотизм	30
трамп и ситуация в стране	30
выборы	28
деньги и работа	26

Положительные и отрицательные

Topic	sentim
коронавирус	31
ложь и страх	31
бог	28
ненависть и обман	28
законопорядок	26
америка, патриотизм	24
человеческие чувства	24
здравоохранение	23
криминал и коррупция	23
расизм	23
трамп и ситуация в стране	23
общественное мнение	22
повседневная жизнь	22
выборы	19
деньги и работа	18

Отрицательные

Topic	sentim
здравоохранение	17
ложь и страх	14
повседневная жизнь	14
бог	12
законопорядок	11
ненависть и обман	10
общественное мнение	10
расизм	10
выборы	9
криминал и коррупция	9
деньги и работа	8
трамп и ситуация в стране	7
человеческие чувства	7
америка, патриотизм	6
коронавирус	3

Положительные

Распределение тональных компонентов в латиноамериканском корпусе

Topic	sentim
демократия	31
венесуэла и россия	29
воровство и терроризм	29
проблемы людей	29
ситуация в мире	25
коммунисты и диктатура	24
ожидания	24
освобождение	24
президент	24
революционные настроения	23
ложь и предательство	22
базовые нужды	20
трамп	19
сша и латинская америка	17
правительство	14

Положительные и отрицательные

Topic	sentim
демократия	24
воровство и терроризм	21
венесуэла и россия	19
проблемы людей	19
коммунисты и диктатура	18
ожидания	17
президент	15
ситуация в мире	15
базовые нужды	14
ложь и предательство	14
революционные настроения	14
освобождение	13
трамп	11
сша и латинская америка	9
правительство	8

Отрицательные

Topic	sentim
освобождение	11
венесуэла и россия	10
проблемы людей	10
ситуация в мире	10
президент	9
революционные настроения	9
воровство и терроризм	8
ложь и предательство	8
сша и латинская америка	8
трамп	8
демократия	7
ожидания	7
базовые нужды	6
коммунисты и диктатура	6
правительство	6

Положительные

Среднее значение тональности для каждой темы

Topic	sentim
бытовые проблемы	-0.220000
жизнь в россии	-0.180000
ситуация в стране	-0.166667
работа	-0.153333
повседневная жизнь	-0.140000
общественные вопросы	-0.133333
образование	-0.106667
нужды человека	-0.100000
коронавирус в мире	-0.093333
споры	-0.093333
удаленные или спам	-0.093333
семья	-0.080000
путин и страна	-0.066667
финансовые проблемы	-0.066667
карантин	-0.053333

Российский корпус

Topic	sentim
коронавирус	-0.186667
америка, патриотизм	-0.120000
ненависть и обман	-0.120000
ложь и страх	-0.113333
человеческие чувства	-0.113333
бог	-0.106667
трамп и ситуация в стране	-0.106667
законопорядок	-0.100000
криминал и коррупция	-0.093333
расизм	-0.086667
общественное мнение	-0.080000
выборы	-0.066667
деньги и работа	-0.066667
повседневная жизнь	-0.053333
здравоохранение	-0.040000

Американский корпус

Topic	sentim
демократия	-0.17
воровство и терроризм	-0.13
коммунисты и диктатура	-0.12
ожидания	-0.10
венесуэла и россия	-0.09
проблемы людей	-0.09
базовые нужды	-0.08
ложь и предательство	-0.06
президент	-0.06
революционные настроения	-0.05
ситуация в мире	-0.05
трамп	-0.03
освобождение	-0.02
правительство	-0.02
сша и латинская америка	-0.01

Латиноамериканский корпус

Тональная шкала тем

Тема	Сумма средних	Среднее суммы
бытовые проблемы	−0.31	−0.15
нужды человека	−0.30	−0.10
ситуация в стране	−0.29	−0.10
общественные вопросы	−0.26	−0.09
коронавирус	−0.23	−0.11
президент	−0.20	−0.10
повседневная жизнь	−0.19	−0.01
ситуация в мире	−0.14	−0.07

Выводы

- Созданы корпусы комментариев к новостным группам на русском, английском и испанском языке с помошью материалов российских, американских и латиноамериканских СМИ;
- С помощью тематического моделирования выделены темы, присутствующие в комментариях, построены тематические профили сообществ;
- Темы размечены при помощи тональных словарей, показаны распределения тональных компонентов в темах;
- Определено среднее значение тональности для каждой темы, на основании которого построена тональная шкала тем для сходных тем из трёх корпусов;
- Вывод исследования: внимание пользователей в комментариях новостных сообществ и каналов всех исследуемых групп сосредоточено больше на общественно-политических темах, нежели чем на личных и бытовых, но бытовые и общественные проблемы находят у пользователей больший эмоциональный отклик, чем темы политики и ситуации в мире.