Проблемы построения корпуса коротких текстов для задачи классификации отзывов на три класса

Юлия Рубцова

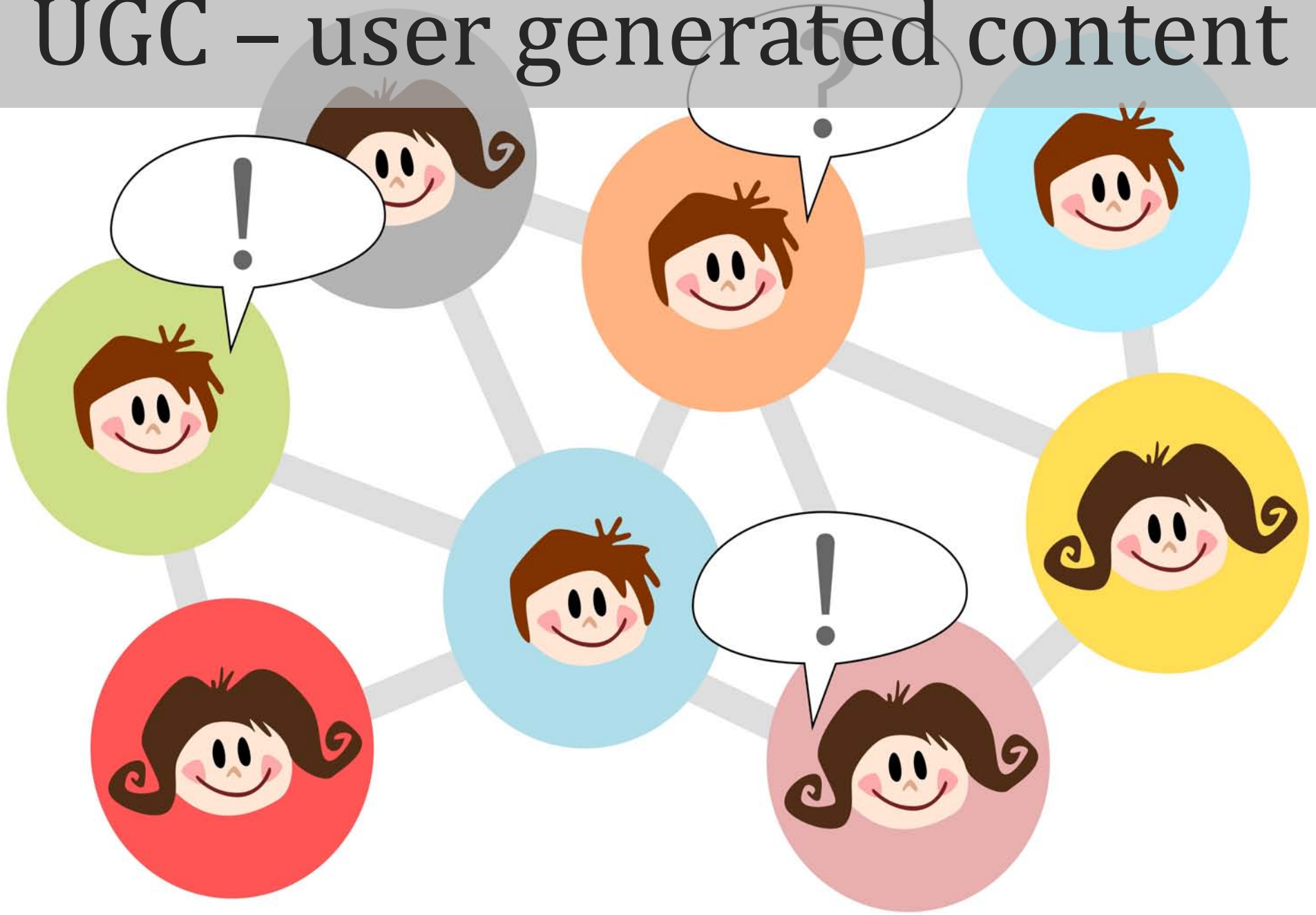
Институт систем информатики
им. А.П. Ершова СО РАН

Содержание

- Обзор предметной области
- Сбор корпуса
- Фильтрация
- Морфологический анализ
- Результаты и закономерности
- Выводы
- Применение

Обзор предметной области

UGC – user generated content



Задачи извлечения и обработки информации из текстов

Прикладные задачи, которые решает автоматическое определение тональности

- исследования отношения потребителей к ее продукции для коммерческой организации;

Apple iPad 4 64Gb Wi-Fi + Cellular

474 оценки

Оцените товар

Сортировать: по дате | по оценке | по полезности

Написать отзыв

Пользователь скрыл свои данные — 23 сентября

отличная модель

Опыт использования: несколько месяцев

Достоинства: Стильный, быстрый, четкий дисплей))) дизайн хорош .качество сборки на высоте!!!

Недостатки: Их мне кажешься нет) единственное это большая цена , но она себя оправдывает)))

Комментарий:
Очень полезная вещ дома на учебе в поездках) пользуюсь им длительное время , очень доволен ! Смотреть фильмы , играть ,серфить интернет одно удовольствие)текст читать с такого экрана очень приятно!!! Не глючит!!!!! Весомые игрушки тянет на ура!!!! Не слушайте людей которые говорят я бы лучше купил 3 планшета за такую цену .... Это не так !!! Когда берешь его в руки сразу чувствуется качество , ни каких скрипов нет!!! Достойная вещ для ценителей удобства , быстроты и качества)))) отзыв написан с него))) все удачи в покупке!!!)))

Отзывы с оценкой

- 37 отзывов
- 13 отзывов
- 3 отзыва
- 3 отзыва
- 3 отзыва

Подробные оценки

Качество изготовления

Адекватность цены

Удобство эксплуатации

Прикладные задачи, которые решает автоматическое определение тональности

- исследования отношения потребителей к ее продукции для коммерческой организации;
- разработка рекомендательной системы для покупателей определенных групп товаров или услуг;

Точность прогноза 2%

0 10 20 30 40 50 60 70 80 90 100

Гарри Поттер и кубок огня

★ ★ ★ ★ ★ ★ ★ ★ ★ ★

Прикладные задачи, которые решает автоматическое определение тональности

- исследования отношения потребителей к ее продукции для коммерческой организации;
- разработка рекомендательной системы для покупателей определенных групп товаров или услуг;
- введение в человеко-машинный интерфейс компьютерной системы, отвечающей за адаптацию поведения системы к текущему эмоциональному состоянию человека

человеко-машинный интерфейс, отвечающей за адаптацию поведения системы к текущему эмоциональному состоянию человека

- психологическое и медицинское диагностирование;
- обеспечение безопасности за счет анализа поведения массовых скоплений людей;
- помощь в проведении оперативно-розыскных мероприятий

Тональность текстов

- Два класса
- Три класса
- 5 классов
- 10 классов
- N-классов

Существующие корпусы текстов

- Корпуса отзывов, содержащие оценки пользователей
- Узкотематические корпуса отзывов (фильмы, книги, техника)
- Корпуса общезначимых новостей (тексты состоящие из нескольких абзацев)

Отличие микроблога от сервиса отзывов

- Обдуманный, структурированный
- Конструктивная критика или похвала
- Не ограничен по длине
- Относится к одной предметной области
- Может одновременно выражать и негатиное отношение и позитивное

- Спонтанны
- Эмоциональны
- Ограничение по длине в 140 символов
- Общетематический ресурс
- Один текст – одна эмоция

Сбор корпуса

Корпус

С высокой точностью можно определить передаваемую эмоцию, если автор указал символ обозначения эмоции на письме (смайлик).
[Метод J.Read 2005]

:) :(:-/ 8-) =(0_o :-D ;-) :-) (((:’(

Подготовка

1. Составлены словари символов, обозначающие на письме:

- отрицательные эмоции,
- Положительные эмоции.

Подготовка

1. Составлены словари символов, обозначающие на письме:

- отрицательные эмоции,
- Положительные эмоции.

2. Сделано допущение, что выражение эмоции относится ко всему сообщению, а не к отдельной его части. (Длина твита 140 символов).

Подготовка корпуса

1. Составлены словари символов, обозначающие на письме:

- отрицательные эмоции,
- Положительные эмоции.

2. Сделано допущение, что выражение эмоции относится ко всему сообщению, а не к отдельной его части. (Длина твита 140 символов).

3. Созданы фильтры для устранения дубликатов и неопределенностей.

Фильтрация

- Положительные и отрицательные эмоции в одном твите
- Retweet
- Копии твитов
- Малоинформативные твиты (<40 символов)
- Реплаи

Сложности

API twitter отдает только 1000 постов на каждый поисковый запрос с географической привязкой

Атрибуты корпуса

- Класс, к которому принадлежит твит (положительный/отрицательный)
- Дата публикации
- Имя автора
- Текст твита
- Количество реплаев
- Количество ретвитов

Тренировочный корпус

- 34235 положительных текстов
- 34225 отрицательных текстов
- 32065 нейтральных текстов

Total: 100 525

Морфологический анализ

Задача

1. Выявить закономерности распределения частей речи между коллекциями заведомо состоящих или не состоящих из эмоционально окрашенных высказываний.
2. Выявить закономерности распределения частей речи между «положительной» и «отрицательной» коллекциями

Инструмент

Tree tagger для русского языка (Shmid 1994), Шаров.

MSD	Animate	Aspect	CATEGORY	Case	Case2	Coord_Type	Definiteness	Degree	Form	Syntactic_Type	Tense	Type	VForm	Voice
Afc			Adjective	-			-	comparative						
Afc---f			Adjective	-			full-art	comparative						
qualificative														
Afp			Adjective	-			-	positive						
Afp---f			Adjective	-			full-art	positive						
qualificative														
Afpfsa			Adjective	accusative			-	positive						
qualificative														
Afpfsaf			Adjective	accusative			full-art	positive						
qualificative														
Afpfsd			Adjective	dative			-	positive						
qualificative														
Afpfsdf			Adjective	dative			full-art	positive						
qualificative														
Afpfs-f			Adjective	-			full-art	positive						
qualificative														
Afpfsg			Adjective	genitive			-	positive						
qualificative														
Afpfsgf			Adjective	genitive			full-art	positive						
qualificative														
Afpfsi			Adjective	instrumental			-	positive						
qualificative														
Afpfsif			Adjective	instrumental			full-art	positive						
qualificative														
Afpfsl			Adjective	locative			-	positive						
qualificative														
Afpfslf			Adjective	locative			full-art	positive						
qualificative														
Afpfsn			Adjective	nominative			-	positive						
qualificative														
Afpfsnf			Adjective	nominative			full-art	positive						

Результаты

Эмоции vs нейтральность

Авторы, выражающие эмоции, склонны использовать

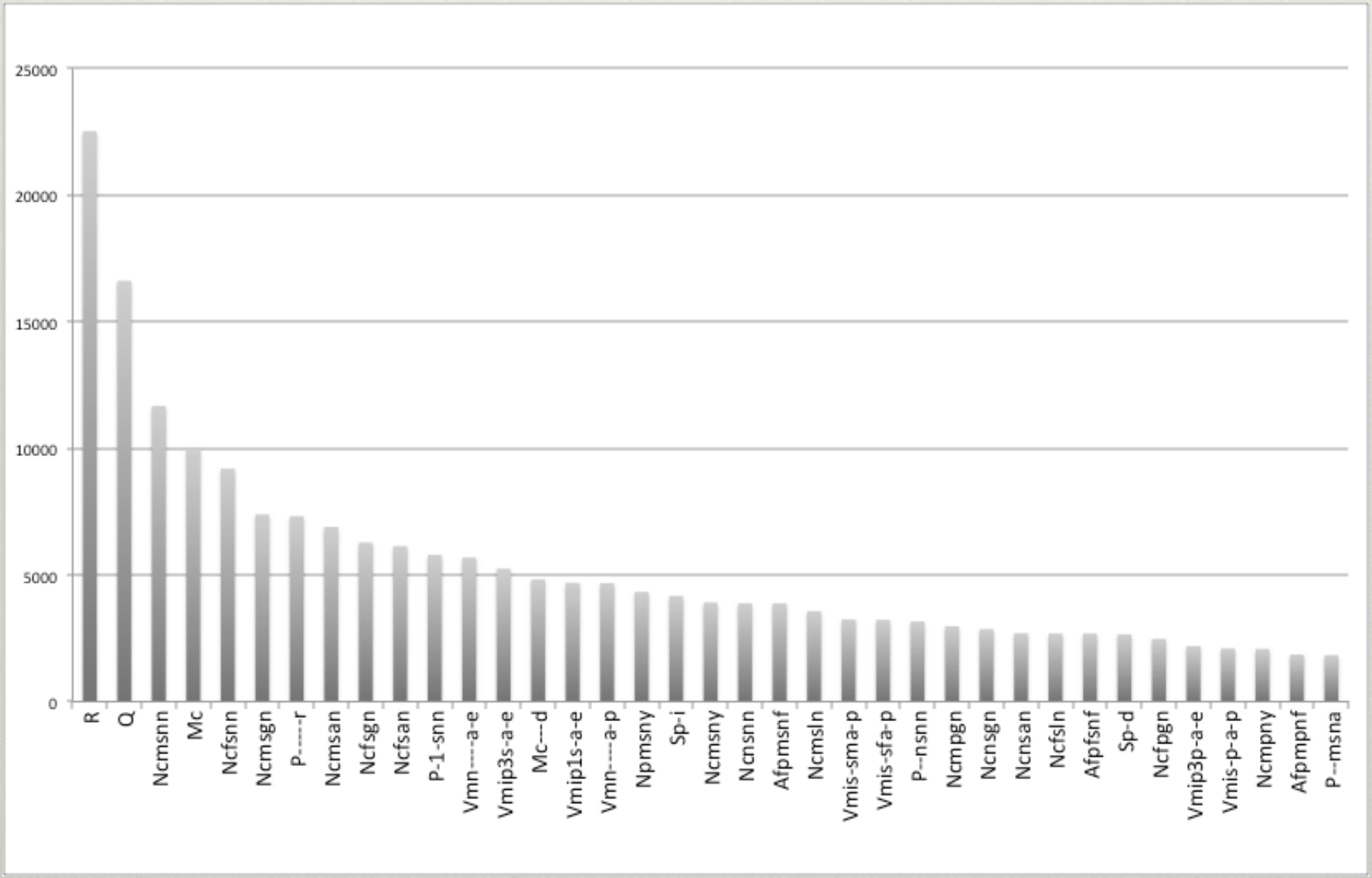
- Наречия
- Частицы
- Существительные и.п. М и Ж рода

Эмоции vs нейтральность

Примеры используемых наречий:

- Прекрасно
- Беспощадно
- Стыдно
- Интересно
- Дико

Эмоции vs нейтральность

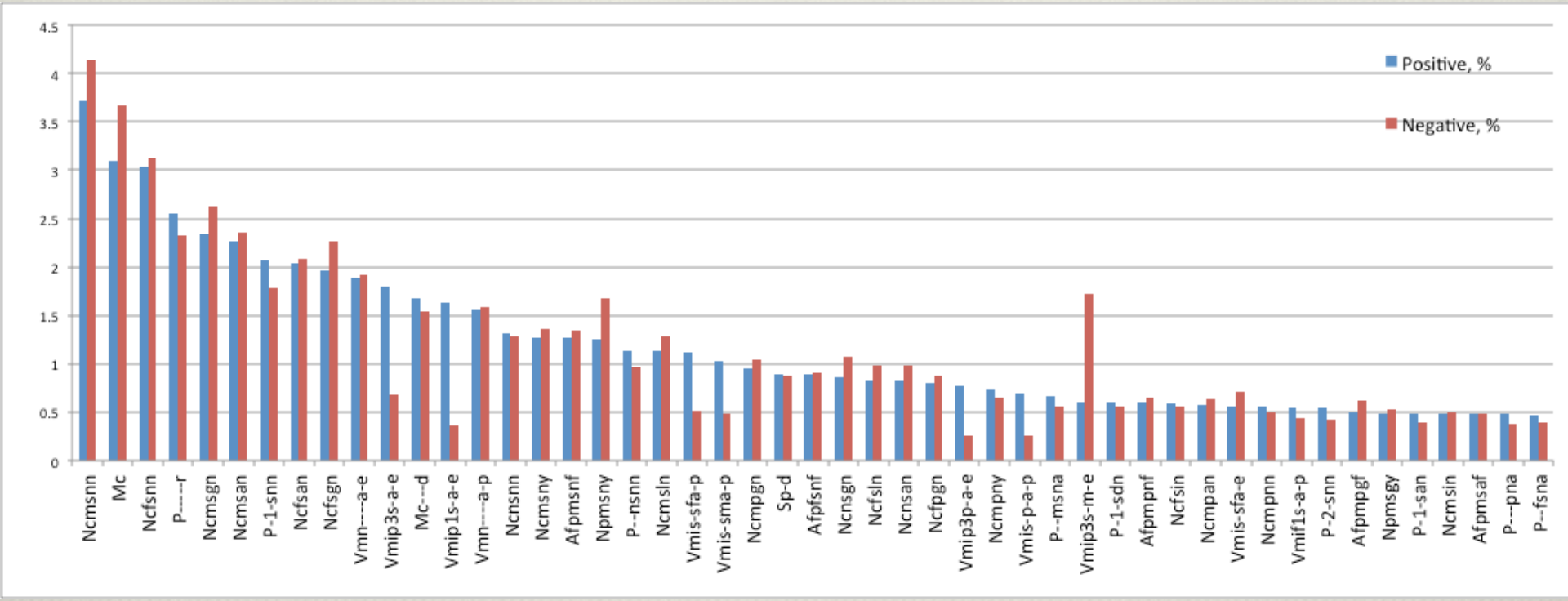


Позитивные VS неативные

- Глагол в активном залоге настоящего времени, 1 и 3 лицо, ед. число.
- Глагол в активном залоге, прошедшее время, 1 и 3 лицо, ед. число.
- Притяжательные местоимения м и ж рода, ед. числе, родительном падеже.

- глаголы настоящего времени, обозначающие продолжительность действия в третьем лице единственного числа несовершенного вида
- сравнительно чаще используются имена собственные в единственном числе в винительном падеже

Позитивные VS неативные



Извлечение оценочных слов из тренировочного корпуса

Словарь стоп-слов

- Фамилии (Навальный)
- Названия продуктов (ФК «Зенит»)
- Яркие мировые события
- Предлоги
- Союзы

Положительные оценочные слова

Вес слова (отношение частоты встречаемости слова в положительных твитах к частоте встречаемости в отрицательных)	Слово	Частота встречаемости слова в коллекции положительных твитов
4.097362044	клип	60
2.230786002	сериал	56
2.22893319	бл*	401
2.192768754	х**	2211
1.657155316	приятный	52
1.6101914	рад	64
1.434076715	зато	60
1.414158983	крутой	71
1.354405787	смеяться	51

Отрицательные оценочные слова

Вес слова (отношение частоты встречаемости слова в отрицательных твитах к частоте встречаемости в положительных)	Слово	Частота встречаемости слова в коллекции отрицательных твитах
13.01650495	продажа	56
10.22725389	утро	484
6.5610792	встать	69
6.136352334	проснуться	88
5.835406825	вставать	53
4.881189357	вчера	119
4.685941782	погибнуть	56
4.623486813	школа	305
4.532532974	рано	52

Выводы

- построен корпус текстов, автоматически размеченный на три класса.
- В корпусе около 100 000 постов
- Каждый текст в корпусе содержит атрибуты, которые помогут сделать выводы об актуальности высказывания и силе его воздействия, важности.
- Корпус морфологически размечен.
- Извлечены оценочных терминов, не относящихся к одной заранее определенной предметной области.

Дальнейшая работа

Тоновый классификатор

Весами для тонового классификатора, на основании которых будет определена вероятность принадлежности высказывания к тому или иному классу являются:

- Актуальность высказывания, сила его воздействия
- Морфология предложений
- Оценочные слова

Разрабатывается программный комплекс для построения корпусов отзывов из разных источников.

Корпуса могут быть использованы для тренировки других классификаторов.

Применение

Тоновый классификатор будет использован для автоматический оценки отзывов на интернет-ресурсы, найденные в качестве кандидатов на включение в интеллектуальные научные интернет-ресурсы (ИНИР)

Спасибо

Вопросы?

Юлия Рубцова