Классификация эмоциональной окраски сообщений в социальных сетях

Н. А. Савинов

Московский физико-технический институт
Факультет Управления и Прикладной Математики
Кафедра Интеллектуальные Системы

Научный руководитель д.ф.-м.н., профессор К. В. Воронцов

Москва,
2013 г.

Постановка задачи анализа тональности

- Дано множество коротких сообщений $C = \{m_j\}_{j=1}^{K}$, относящихся к компании и ее продуктам.
- Классифицировать C на три группы сообщений:
  1. тональные отрицательные
  2. тональные положительные
  3. нейтральные
- Правильную классификацию на обучающей выборке определяет эксперт-асессор.
- Функционалы качества: точность P, полнота R, F-мера относительно классов тональных, положительных и отрицательных сообщений.

Эталонный метод

Два этапа:

- Э1: классификация тональный/нейтральный
- Э2: классификация положительный/отрицательный

Метод классификации — логистическая регрессия
Признаки — частоты внутрисловных 4-грамм

L. Barbosa, J. Feng “Robust sentiment detection on Twitter from biased and noisy data”. International Conference on Computational Linguistics. 2010.

Эксперимент 1 с эталонным методом

Данные:

- Э1: 26000 размеченных сообщений из Твиттера про компанию Яндекс
- Э2: 600000 отзывов о товарах на Яндекс.Маркете с оценками $\underbrace{1, 2}_{-}, 3, \underbrace{4, 5}_{+}$

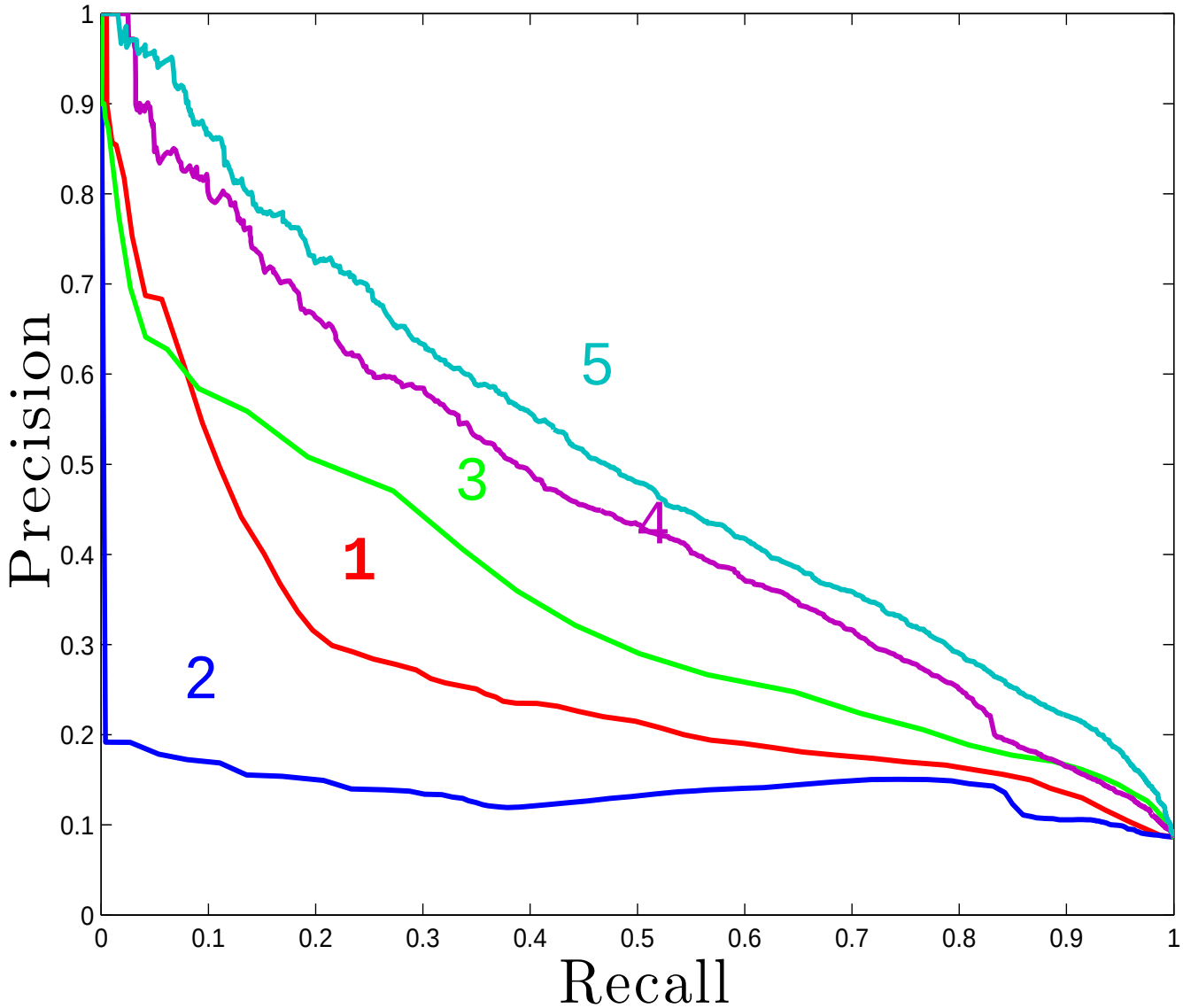
Результат:

Тип Э1	$F^{pos/neg}$	F^{tonal}	R^{tonal}	R^{pos}	R^{neg}	P^{tonal}	P^{pos}	P^{neg}
Идеальный	0,719	1	1	0,646	0,789	1	0,654	0,786
Реальный	0,345	0,441	0,901	0,464	0,808	0,239	0,282	0,158

- При идеальном Э1 прирост $\Delta F^{pos/neg} = 0,374$
- При идеальном Э2 прирост $\Delta F^{pos/neg} = 0,281$

Вывод: Улучшение качества Э1 — более важная задача.

Эксперимент 2: последовательное улучшение метода



Эксперимент 2: шаги улучшения

1. Эталонный метод — логистическая регрессия для Э1 и Э2
2. Бутстрэп — учет несбалансированности выборки
3. Вероятностный бутстрэп — учет полудублей
4. Признаки на основе синтаксиса, морфологии и меток тональности слов
5. Признаки на основе мета-информации Твиттера

Учет несбалансированности выборки: метод бутстрэпа

Вход: двухклассовая выборка S, мощности классов m и $n \gg m$;

Выход: сбалансированная выборка E из $2m$ объектов;

1: Выборка $E := \{m$ объектов из класса-меньшинства и m случайных объектов из класса-большинства$\}$;
2: **повторять итерации:**
3: На выборке E обучить классификатор;
4: Полученный классификатор применить ко всем объектам из класса-большинства, для каждого объекта получить вещественную оценку — вероятность принадлежать классу-меньшинству;
5: Выборка $E := \{m$ объектов из класса-меньшинства и m объектов класса-большинства с наибольшей вероятностью ошибки$\}$;

Вероятностный бутстрэп

Проблема: обычный бутстрэп не работает из-за большой доли полудублей.

Решение:

- Пусть для каждого объекта x_i нам известна вероятность неправильной классификации p_i^{error}.
- Введем на объектах класса-большинства распределение $P(i) \sim \exp^{-\frac{1-p_i^{error}}{T}}$, где T — параметр температуры, отвечающий за "остроту" пиков распределения.
- Подвыборка из m объектов класса-большинства порождается случайно из распределения $P(i)$.
- Перебором по сетке найдено оптимальное значение $T = 3$.

Учет синтаксиса, морфологии и меток тональности слов

Проблемы эталонного метода:

- 4-граммы приводят к появлению ложных тональных слов.
- Не учитывается взаимное расположение слов и синтаксические связи.
- Не учитывается тональность, присущая отдельным словам.
- Не учитываются эмотиконы (например, ":)").

Решение:

- Вместо 4-грамм использовать лемматизацию.
- Использовать биграммы и трехграммы слов по последовательности и синтаксическому дереву.
- Использовать метки частей речи, метки эмотиконов и метки тональности слов.

Учет синтаксиса, морфологии и меток тональности слов: особенности подхода

Униграммный подход:

яндекс	очень	помог

Униграммы: яндекс, очень, помог.

Подход на основе синтаксиса, морфологии и меток тональности слов:

яндекс сущ.	очень нареч.	помогать глаг. positive

К униграммам **добавляются:**

Биграммы по синтаксическому дереву: помогать $\rightarrow$ яндекс, positive $\rightarrow$ яндекс, глаг. $\rightarrow$ яндекс, ...

Биграммы по последовательности: очень_помогать, очень_positive, очень_глаг., ...

Трехграммы по синтаксическому дереву:

яндекс $\leftarrow$ помогать $\rightarrow$ очень, яндекс $\leftarrow$ positive $\rightarrow$ очень, ...

Трехграммы по последовательности: яндекс_очень_помогать, яндекс_очень_positive, ...

Признаки на основе мета-информации Твиттера

- Ранее рассматривалась только текстовая информация, содержащаяся в сообщении.
- Предлагается использовать следующую информацию, предоставляемую Твиттером:
  1. Имя пользователя.
  2. Наличие ретвита и наличие непустого ретвита.
  3. Является ли внешняя ссылка файлом или путем (заканчивается на знак “/” или на домене).
  4. Наличие домена yandex во внешней ссылке.
  5. Ключевые слова, выделенные тэгом <b> в выдаче поиска по блогам.
- Признаки кодируются специальными ключевыми словами.

Недостатки предложенного алгоритма (1)

- Недостаточный размер выборки для подзадачи нейтральный/тональный. При размере выборки 26000, среди них только 2600 сообщений являются тональными.
- Отсутствие масштабной разметки слов по тональности. В данной работе применяется словарь размером примерно 2000 слов, и этого явно недостаточно. Реальные примеры:
  - “Я смотрю Yandex Fotki сильно допилили за год. Думаю, может туда переехать с Flickr. Хм?”. 33% вероятность быть тональным.
  - “ Моё доверие яндекс утратил...Как так можно...Перепутать Россия 2 и Первый? Когда Россия-Канада играют... ”. Здесь выражается отрицательная тональность, но алгоритм присваивает лишь 55% вероятность быть тональным.

Недостатки предложенного алгоритма (2)

- Недостаточно качественное выделение объекта оценки. Пример реального сообщения: “закончится эта неделя, закончится мой круг ада, тобеш закончатся все контрольные и зачеты”. Отнесен к тональным с вероятностью 82%.
- Сложность машинного анализа смысла естественного языка. Пример: “не нравится мне, что в яндекс-картах весь литейный красный;(((”. Это сообщение является тональным с вероятностью 90%.

Результаты

- Разработан 2-х этапный метод классификации эмоциональной окраски сообщений.
- Предложен метод вероятностного бутстрэпа для несбалансированных выборок с полудублями.
- Предложен метод учета дополнительной информации о морфологии, синтаксисе, метках тональности слов и мета-информации из Твиттера.
- Показано, что в совокупности применение предложенных методов позволяет улучшить точность и полноту.