Предсказание интересов пользователей социальных сетей по текстовым сообщениям

Выполнил:
Николаев Кирилл Игоревич,
Студент группы 19ИАД,
НИУ ВШЭ - НН

Научный руководитель:
Доктор технических наук,
Профессор кафедры ИСиТ,
Савченко Андрей Владимирович

Нижний Новгород, 2020

Определение пользовательских интересов

- Рост популярности социальных сетей
  - Анализ пользовательских данных: рекомендательные системы, целевая реклама;

- Большинство использует личные данные (возраст, национальность) или метаданные (история покупок, оценки);

- Актуально применить автоматическую обработку текстов: решить задачу автоматического определения интересов и предпочтений пользователя по содержанию написанных им текстов.

Объект и предмет исследования

- Объект — определение пользовательских интересов;
- Предмет — предсказание интересов пользователей по текстовым сообщениям;

Цель и задачи исследования

- Цель исследования: разработка алгоритма классификации интересов пользователя по набору написанных им текстов;
- Задачи:
  - Провести аналитический обзор современной литературы по существующим подходам к задаче;
  - Осуществить сбор текстовых данных для решения данной задачи;
  - Разработать алгоритм идентификации предпочтений пользователя по набору написанных им текстов;
  - Провести экспериментальные исследования различных подходов к решению задачи.

Формальная постановка задачи классификации текстов

- Набор докментов $D = \{d_1, \ldots, d_n\}$;
- Конечный набор классов $C = \{c_1, \ldots, c_m\}$ — выраженные в тексте интересы:
  - Каждому d соответствует как минимум один c;
- Найти функцию f :
  - Для любой пары $<d, c>$ определить, соответствует ли документ интересу: $f: D \times C \to \{0, 1\}$

Два способа определения интересов

- Классификация одновременно целого корпуса текстов одного пользователя:
  - Каждый документ d_i — целая совокупность текстов, каждому ставится в соответствие не менее 1 и не более m классов c. Решается задача многоклассовой классификации;

- Последовательная классификация каждого из текстов с последующим объединением результатов классификации:
  - Каждый документ принадлежит всего одному классу и уникальному пользователю. Сведение задачи к множеству решений одноклассовой классификации, совокупность результатов формирует профиль интересов.
- В ходе исследования мы использовали второй подход.

Существующие методы классификации текстов

- Классические: наивный байесовский классификатор, логистическая регрессия, метод опорных векторов, случайный лес (реализовывались с Scikit-learn);
- Традиционные методы глубинного обучения - рекуррентные и свёрточные нейронные сети (реализовывались с TF 2.0):
  - Рекуррентные — фиксируют предыдущие состояния, используя, таким образом, контекст. Недостаток — длительное время работы;
  - Свёрточные — фильтр позволяет использовать матричное представление предложений для эффективной классификации.

Существующие методы классификации текстов

- Новейшие методы:
  - ELMO [Peters et al., 2017]. Первой использовала внутренние состояния глубокой двунаправленной языковой модели для учёта глубокой семантики и преодоления SOTA в разных задачах;
  - Меньше, чем через год: BERT [Devlin et al., 2018]. Логическое развитие ELMO. Т.н. Masked Language Model. Часть токенов маскируется, задача сети — определить их, используя и левый, и правый контексты. SOTA в различных задачах обработки текста и по сей день.

Существующие методы векторного представления

- Традиционные методы:
  - Мешок слов;
  - Словесные и символьные N-граммы;
  - Двоичные признаки;
  - Признаки на основе регулярных выражений;
- Распределённые представления:
  - Word2vec, Doc2vec, FastText.
- Эмбеддинги имеют меньшую размерность и моделируют семантику, однако не всегда интерпретируемы.

Набор данных

- Отличие от существующих датасетов, таких, как Taiga: https://tatianashavrina.github.io/taiga_site/)
  - Только тексты соцсетей (форумы);
  - Короткие тексты;
  - Taiga: 2% соцсети vs 77% литературных текстов;

Набор данных

- 239 089 текстов:
  - Форумные собщения: forum.kinopoisk.ru
- Десять классов: аниме, искусство, книги, еда, фильмы, футбол, игры, музыка, природа, путешествия;
- Среднее количество слов по корпусу — 30; более крупные тексты разбиты с сохранением класса и авторства: 376 270 текстов;
- 54.1% > 150 символов: 189 823 текста;
- Очень несбалансированно;
- Тренировочная, тестовая выборки:
  - 5% от объёма корпуса с сохранением распределения;

Распределение по классам

Категория	До фильтрации		После разделения		После фильтрации	
Аниме	8466	3.54%	12297	3.27%	5765	3.04%
Еда	13275	5.55%	22340	5.94%	12495	6.58%
Живопись	2806	1.17%	5504	1.46%	3445	1.81%
Игры	77454	32.40%	122806	32.64%	62749	33.06%
Книги	19064	7.97%	40215	10.69%	26901	14.17%
Музыка	27819	11.64%	39358	10.46%	16925	8.92%
Природа	3285	1.37%	4644	1.23%	2041	1.08%
Путешествия	3327	1.39%	7577	2.01%	5230	2.76%
Фильмы	14186	5.93%	22884	6.08%	11856	6.25%
Футбол	69407	29.03%	98645	26.22%	42416	22.35%
Итого	239089		376270		189823	

Предобработка

- Стоп-слова, латинница, лемматизация RNNMorph.
- До:
  - **Я кофе только со сливками пью.А чай я пью то-же только горячий если даже пару минут после кипения прошло,я снова его включаю:Кстати врач сказал,что такой горячий нельзя пить,но это уже бесполезно я уже зависим**
- After:
  - **кофе сливка пить чай пить горячий пара минута кипение пройти снова включать кстати врач сказать горячий пить это бесполезный зависеть**

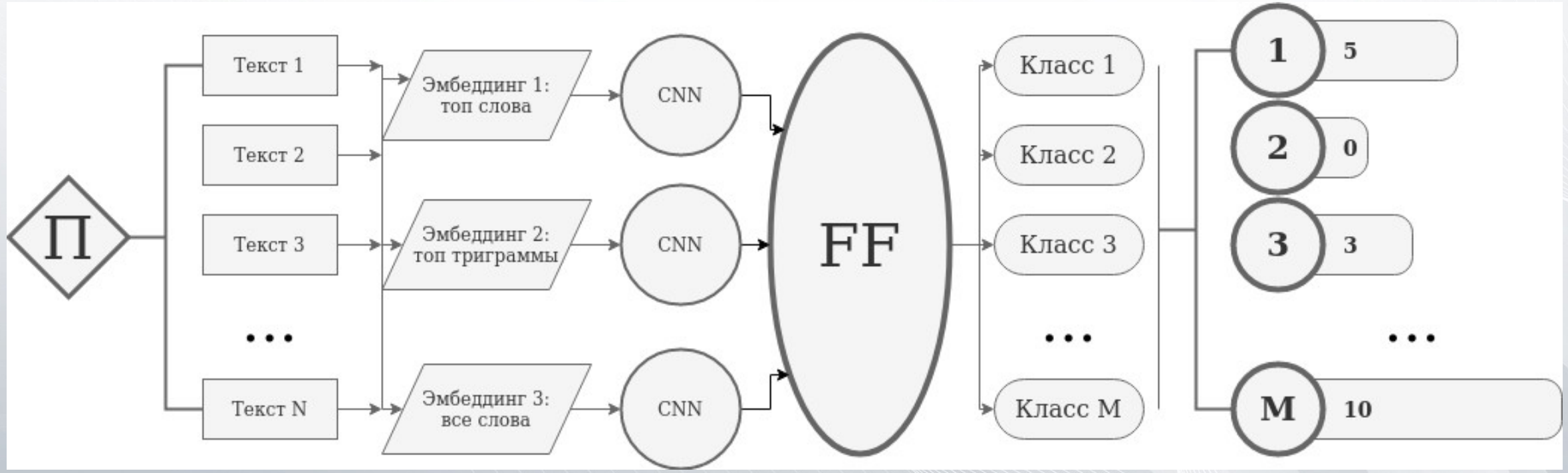
Представление текстов

- Doc2Vec: 300d, 600d;
- 10 x2 сложных признаков:
  - Топ слов по PPMI ;
  - Топ символьных триграмм по PPMI (поточечная взаимная информация).
- Пример:
  - Еда: *аппетит, кофе, блюдо, гарнир; 'ыр ', 'кеф', 'ощ ', 'сыр'*
  - Книги: *слог, паланик, книжный, роман; 'афк', 'гюг', 'фка', 'руэ', 'дюм', 'элш', 'амю', 'юма'*.
  - Футбол: *поражение, юве, полуфинал, счёт; 'уеф', 'лч', 'уцк', 'цс'*

Поточечная взаимная информация [Bouma, 2009]

- Какие слова и символьные триграммы устойчиво встречаются в текстах одного типа и мало в других;
- Значения признаков: пропорция специфичных для класса элемента относительно всех слов / триграмм в тексте;
- 20 значений: 10 для слов, 10 для симв. триграмм каждого класса.

Предложенный подход



Пример применения алгоритма

- Пользователь: Sunnywolf
- Сообщения:
  - Да, я считаю что Тетрадь Смерти — лучшее что случалось с аниме. Я много тайтлов видел, и более популярного не было ни до ни после. Кто бы мог подумать. *(Аниме);*
  - Я не думаю, что оскар достанется Дикаприо. Да, Тарантино написал хороший сюжет, и фильм было очень интересно смотреть, но вы только посмотрите на других кандидатов в этом году. Фантастика. *(Фильмы)*

Пример применения алгоритма

- Пользователь: Sunnywolf
- Сообщения:
  - А ты что думаешь про работы Миядзаки? Моя любимая- «Унесённые призраками», но и Порко Россо недалеко. У него буквально нет плохих работ. *(Аниме)*
  - А ты правда думал, что наши победят? Они с Испанией и так выше головы прыгнули, АКинфеев красава, не спорю, но полуфинал — предел, дальше они не попадут. *(Футбол)*
  - Я думаю они попадут на следующие International, слишком сильная команда. Киберспорт будет уже не таким, когда они уйдут. *(Игры)*.

Примерный результат работы

Категория	Всего	Есть интерес?
Аниме	2	1
Еда	0	0
Живопись	0	0
Игры	1	1
Книги	0	0
Музыка	0	0
Природа	0	0
Путешествия	0	0
Фильмы	1	1
Футбол	1	1

Традиционные методы (10+10)

Model	Accuracy	Точность	Полнота	F-мера
Random Forest Classifier – 100 e.	0.56	0.56	**0.72**	0.55
Multinomial Naïve Bayes	0.64	0.64	0.66	0.65
LinearSVC	0.65	0.65	0.65	0.64
Voting Classifier (the 3 above)	**0.66**	**0.66**	0.66	**0.66**
Random Forest Classifier – 100 e.	0.56	0.56	**0.72**	0.55

Традиционные методы (BoW, 100)

Model	Accuracy	Точность	Полнота	F-мера
ComplementNB	**0.71**	**0.71**	0.75	0.70
LinearSVC	**0.71**	**0.71**	**0.80**	**0.73**
Random Forest Classifier	0.61	0.61	0.74	0.62
Voting Classifier (gauss – linsvc – rfc 100)	0.70	0.70	0.76	0.71
ComplementNB	**0.71**	**0.71**	0.75	0.70

Глубинное обучение на комбинациях признаков

Модель	Валидационная выборка	Тестовая выборка
300, d2v: Bi-LSTM 100 – Dense 200 – Dense 100 – d/o 0.2	0.67	0.669
310, d2v-words	0.767	0.745
310, d2v-trigrams	0.756	0.722
320, d2v-words-trigrams: Bi-LSTM 100 – Dense 100, d/o 0.2	**0.796**	**0.785**

FastText и Bert

Модель	Точность	Время обучения (ч.)
BERT (полная предобработка)	0.82	~4
FastText	0.85	~6
BERT (без членения текстов)	0.95	~4

CNN + эмбеддинг-слои

Архитектура	Accuracy	Точность	Полнота	F-мера
((Embedding → CNN → Bidirectional LSTM) + (Feedforward 20 - 10)) → Dense 200 - 100	0.816	0.82	0.82	0.81
BERT (сокращённые тексты)	0.82	-	-	-
((Embedding → Bidirectional LSTM) + (Feedforward 20 - 10)) → Dense 200 - 100	0.837	0.84	0.84	0.83
((Embedding → CNN) + (Feedforward 20 – 10)) → Dense 200 - 100	0.839	0.84	0.84	0.83
FastText	0.85	0.85	0.85	0.84

CNN + эмбеддинг-слои

Архитектура	Accuracy	Точность	Полнота	F-мера
((Top Trigram Embedding → CNN) + (Top Word Embedding → CNN) + (All Words Embedding → CNN)) → Dense 512 — 256: Conv1D, не сокращённые тексты	0.88	0.89	0.88	0.87
((Top Trigram Embedding → CNN) + (Top Word Embedding → CNN) + (All Words Embedding → CNN)) → Dense 512 — 256: Conv1D, сокращённые тексты	0.93	0.93	0.93	0.93
BERT (не сокращённые тексты)	0.95	-	-	-

Матрица ошибок для Conv1D

	Anime	Art	Books	Films	Food	Football	Games	Music	Nature	Travel
Anime	212	3	10	8	2	0	43	3	1	0
Art	3	139	25	6	1	5	3	4	5	5
Books	5	10	1175	12	12	6	50	13	11	8
Films	3	7	29	532	3	3	23	8	0	3
Food	1	0	15	1	587	1	14	2	4	9
Football	2	1	5	4	1	2049	37	6	2	15
Games	6	1	18	14	6	12	3089	5	4	9
Music	2	2	18	6	2	3	30	736	2	6
Nature	0	2	4	1	5	0	12	1	84	9
Travel	0	0	6	4	6	8	8	3	11	210

Статистика Conv1D

	Точность	Полнота	F-мера	Support
Anime	0.91	0.75	0.82	282
Art	0.84	0.71	0.77	196
Books	0.90	0.90	0.90	1302
Films	0.90	0.87	0.89	611
Food	0.94	0.93	0.93	634
Football	0.98	0.97	0.97	2122
Games	0.93	0.98	0.95	3164
Music	0.94	0.91	0.93	807
Nature	0.68	0.71	0.69	118
Travel	0.77	0.82	0.79	256
Accuracy			0.93	
Weighted avg	0.93	0.93	0.93	9492

Итог

- Сформулирована задача автоматического определения интересов пользователей по текстовым сообщениям на русском языке;
- Собран новый корпус и размещён в открытом доступе: https://github.com/Pythonimous/forum-classifier);
- Выбраны наиболее эффективные технологии классификации текстов;
- Разработан и предложен алгоритм определения пользовательских интересов по текстовым сообщениям;

Итог

- Разработаны комплексные признаки: информативные по PPMI элементы;
- Наилучшая точность достигается с помощью BERT;
- Эмбеддинги + свёрточная сеть — сравнимые показатели:
  - Хотя точность на 2% хуже топового результата, обучение и статистический вывод происходят гораздо быстрее: лучше подходит для реальной задачи.

Пути улучшения

- Увеличение объёма данных, особенно для малопредставленных классов;
- Более комплексные архитектуры и представления;
- На следующем этапе исследования — подход с полноценной multi-label моделью и решение более сложной задачи извлечения предпочтений из неструктурированного текста (т. е. фактически объектно-ориентированного анализа тональности).

Предсказание интересов пользователей социальных сетей по текстовым сообщениям

Выполнил:
Николаев Кирилл Игоревич,
Студент группы 19ИАД,
НИУ ВШЭ - НН

Научный руководитель:
Доктор технических наук,
Профессор кафедры ИСиТ,
Савченко Андрей Владимирович

Нижний Новгород, 2020