Анализ тональности русскоязычных текстов при помощи рекуррентных нейронных сетей с механизмом внимания

Илья Сергеевич Иванов
Научный руководитель к.ф.-м.н. Михаил Бурцев

Московский физико-технический институт
Факультет Инноваций и Высоких Технологий
Кафедра Анализа Данных

2017

Цель исследования

Исследовать новые методы анализа тональности коротких текстов на русском языке с применением рекуррентных нейронных сетей и механизма внимания.

Проблемы

Сложная морфология русского языка.
Особенности лексикона пользователей соц. сети.
Малый объём данных для обучения.

Предположения

Зависимость класса от порядка слов в тексте.
Разная значимость слов в тексте при классификации.

Литература

1. Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D.. Comparison of Neural Network Architectures for Sentiment Analysis of Russian Tweets. Computational Linguistics and Intellectual Technologies. Dialog, 2016.
2. Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, Eduard H. Hovy. Hierarchical Attention Networks for Document Classification. HLT-NAACL, 2016.
3. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2014.

Постановка задачи классификации

Дано множество коротких сообщений $\mathfrak{D} = \{\mathbf{d}_j\}_{j=1}^{K}$, относящихся к компании(-ям).

Необходимо классифицировать сообщения из $\mathfrak{D}$ на три класса:

1. положительной тональности (положительные);
2. отрицательной тональности (отрицательные);
3. не имеющие тональности (нейтральные).

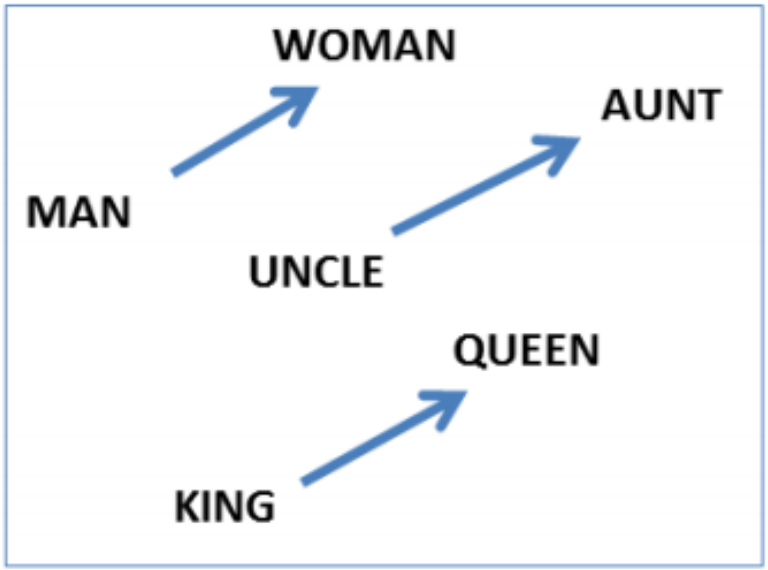
Функционалы качества

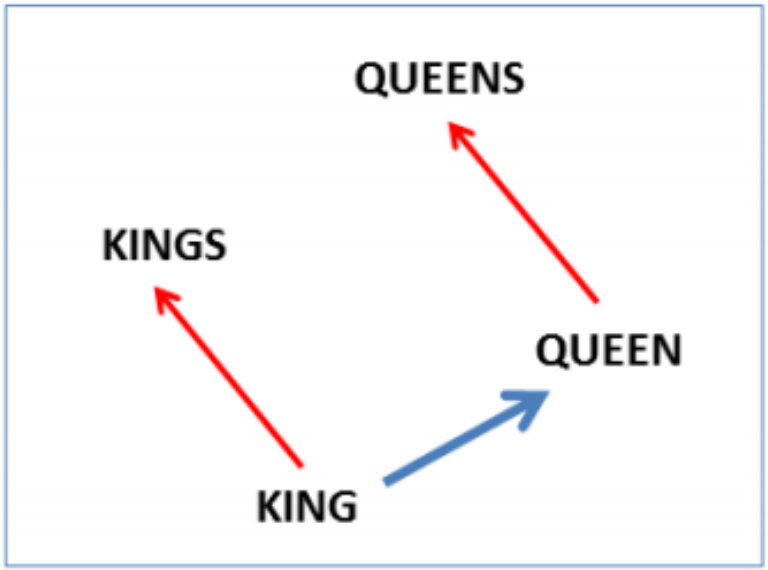
Макро-усредненная F-мера относительно классов положительных и отрицательных сообщений.

В качестве классификатора предлагается использовать двунаправленную рекуррентную нейронную сеть с механизмом внимания.

Векторное представление слов

- Сообщение $\mathbf{d} \in \mathfrak{D}$ является последовательностью слов $\mathbf{d} = \mathbf{w}_1..\mathbf{w}_T$ из словаря $\mathfrak{W}$.
- Слово $\mathbf{w} \in \mathfrak{W}$ представляется вектором в D-мерном пространстве.
- Векторное представление для всех слов из словаря получается при помощи алгоритма Word2Vec, применённом на большом наборе неразмеченных данных.





(Mikolov et al., NAACL HLT, 2013)

Рекуррентная нейронная сеть

- В качестве классификатора используется двунаправленная рекуррентная нейронная сеть типа GRU (Gated Recurrent Unit) с механизмом внимания.
- Функцией ошибки является перекрёстная энтропия для трёх классов.

$$J(W) = -\sum_{i=1}^{n}\sum_{k=1}^{3} y_i^{(k)} \log \hat{y}_i^{(k)},$$

$$\hat{y}_i^{(k)} = \frac{\exp s_i^{(k)}}{\sum_{j=1}^{3} \exp s_i^{(j)}}$$

Двунаправленный GRU

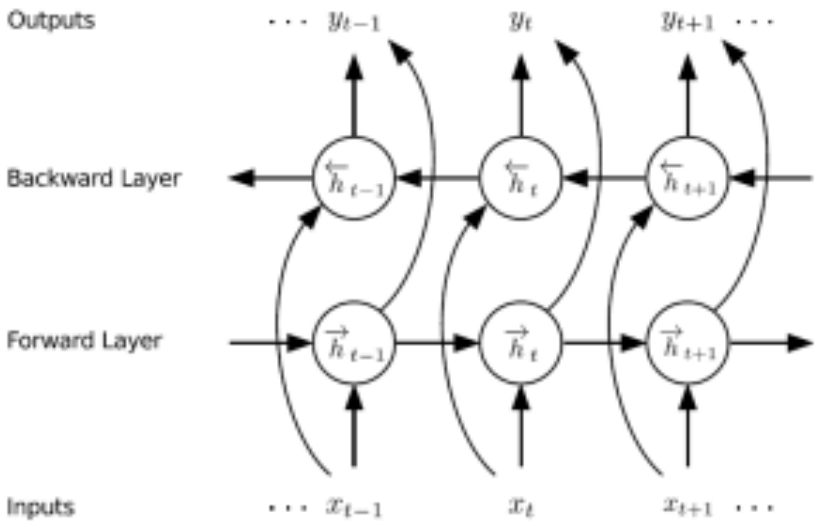
Уравнения GRU

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1}) \quad (1)$$

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1}) \quad (2)$$

$$\tilde{h}_t = \tanh(W x_t + U(r_t \circ h_{t-1})) \quad (3)$$

$$h_t = (1 - z_t) \circ \tilde{h}_t + z_t \circ h_{t-1} \quad (4)$$

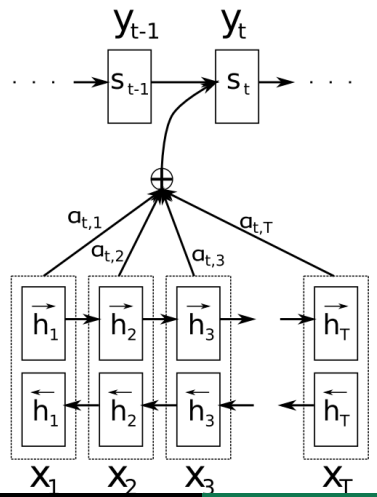


Уравнения механизма внимания

$$v_t = \tanh\left(W_\omega \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] + b_\omega\right) \tag{5}$$

$$\alpha_t = \frac{\exp\left(v_t^T u_\omega\right)}{\sum_{j=1}^{T} \exp\left(v_j^T u_\omega\right)} \tag{6}$$

$$v = \sum_{t=1}^{T} \alpha_t h_t \tag{7}$$



Набор данных

В качестве коллекции документов $\mathfrak{D}$ используется набор сообщений пользователей соц. сети Twitter с упоминанием некоторых банков и телекоммуникационных компаний, собираемые с 2013-го года. Особенностями данной коллекции являются:

- Размер сообщения - не более 140 символов
- Лексикон:
  - сленг
  - сокращения
  - эмотиконы
- Спец. символы:
  - # (хэштег)
  - @ (ссылка на пользователя)
- Ссылки на внешние ресурсы

Цели эксперимента

1. Реализовать архитектуру двунаправленной рекуррентной сети с механизмом внимания (Python + TensorFlow)
2. Провести подбор оптимальных гиперпараметров (GridSearch)
3. Сравнить результаты с предложенными ранее алгоритмами.

Вычислительный эксперимент

В ходе эксперимента сравниваются результаты предложенного алгоритма классификации с такими алгоритмами как двунаправленная рекуррентная нейронная сеть (без механизма внимания) и метод опорных векторов.

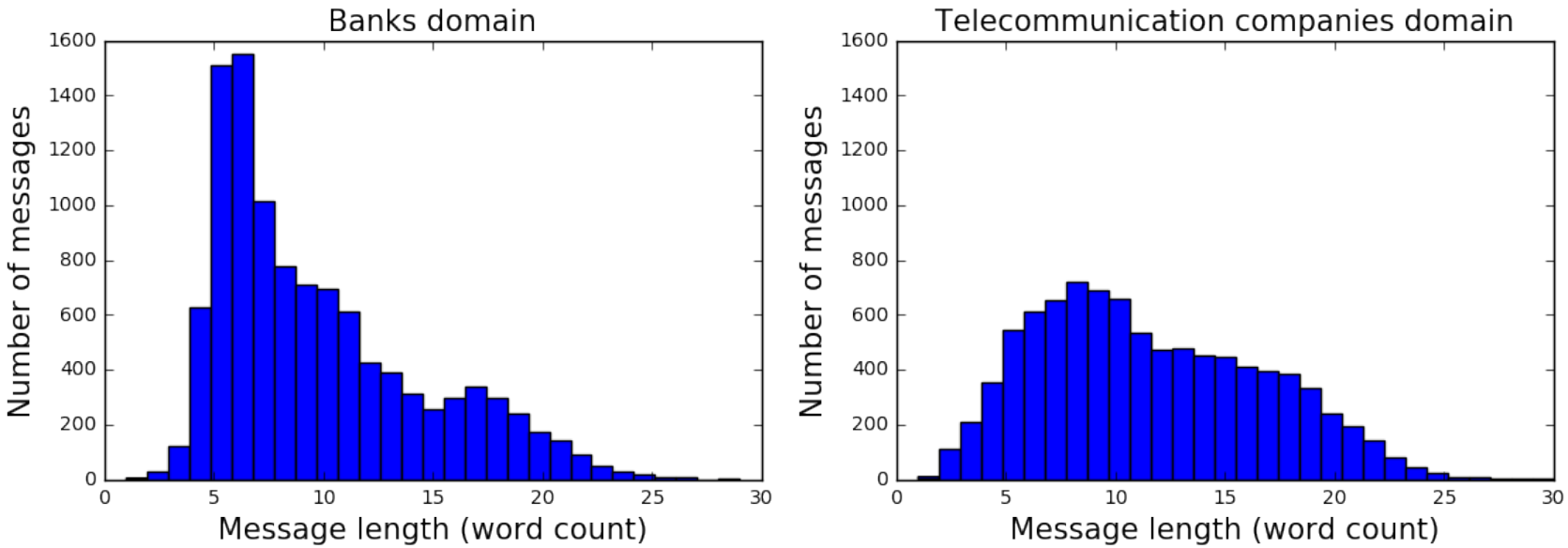
План эксперимента

1. Предобработать набор текстов
2. Обучить Word2Vec
3. Реализовать двунаправленный GRU
4. Реализовать механизм внимания
5. Подобрать оптимальные параметры модели на обучающей выборке
6. Протестировать модель на отложенной выборке
7. Сравнить результаты с другими алгоритмами

Предобработка данных

1. Токенизация (NLTK)
2. Лемматизация (PyMorphy2)
3. Векторизация слов (Word2Vec, обученный на русскоязычном корпусе из социальных медиа)
4. Дополнение последовательностей нулями до максимальной длины (zero-padding)

Рис.: Распределение кол-ва слов в сообщении



Сравнение качества полученных моделей

Таблица: Результаты моделей на 5-фолд кросс-валидации

	Banks	Telecommunication companies
	F1-score	F1-score
Bi-GRU	0.740	0.625
Bi-GRU + Attention	0.737	0.609
2-layer GRU, reversed sequences (Arhipenko)	0.621	0.660
Bi-GRU (Arhipenko)	0.621	0.652
LSTM (Arhipenko)	0.603	0.641

Сравнение качества полученных моделей

Таблица: Результаты моделей на тестовой выборке

	Banks	Telecommunication companies
	F1-score	F1-score
Bi-GRU	0.48	0.52
Bi-GRU + Attention	0.51	0.49
2-layer GRU, reversed sequences (Arhipenko)	0.55	0.56
CNN (Arhipenko)	0.48	0.47
SVM baseline	0.46	0.46
Majority baseline	0.31	0.19

Заключение

- Реализован алгоритм двунаправленной рекуррентной нейронной сети с механизмом внимания для классификации тональности коротких русскоязычных текстов. Код отлажен и выложен в открытый доступ
- Проведён поиск оптимальных гиперпараметров алгоритма
- Проведено сравнение результатов с предложенными ранее алгоритмами
- Подготовлен отчет по результатам работы

Дальнейшее исследование

Исследование применимости данной модели в качестве модуля для нейронной сети, генерирующей сообщения с заданной тональностью.