УДК 004. 934.2

Магистрант А. С. Демещик
Науч. рук. проф. И. Г. Сухорукова
(кафедра программной инженерии, БГТУ)

ПОДХОДЫ К ПРОВЕДЕНИЮ СЕНТИМЕНТ-АНАЛИЗА

Сентимент-анализ – класс методов контент-анализа в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов (мнений) по отношению к объектам, о которых говорится в тексте.

Тональность – это эмоциональное отношение автора высказывания к некоторому объекту (объекту реального мира, событию, процессу или их свойствам/атрибутам), выраженное в тексте. Эмоциональная составляющая, выраженная на уровне лексемы или коммуникативного фрагмента, называется лексической тональностью (или лексическим сентиментом). Тональность всего текста в целом можно определить как функцию (в простейшем случае сумму) лексических тональностей составляющих его единиц (предложений) и правил их сочетания.

Для цельного понимания сентимент-анализа необходимо сформулировать его главную задачу, которая вкратце звучит следующим образом: нахождение в текстовой информации мнений и выявление их свойств. Мнение же, в свою очередь, может быть определено кортежем из 5 элементов (e, f, op, h, t):

- e – entity – объект мнения;
- f – feature – свойств(о/а) объекта;
- op – orientation/polarity – тональная оценка мнения;
- h – holder – субъект(владелец) мнения;
- t – time – момент времени, в который было выражено мнение.

Приведем пример. Ниже представлен комментарий, размещенный на форуме:

Джон Х., 05.02.2019: «Я вчера купил новую фотокамеру и она превосходна!»

Выделим элементы мнения:

- e – entity – фотокамера;
- f – feature – обобщение;
- op – orientation/polarity – положительная;
- h – holder– Джон Х.;
- t – time – 05.02.2019.

Полученный кортеж полностью характеризует ситуацию, описанную в комментарии. Мы знаем объект, относительно которого выражается мнение, его эмоциональную окраску и человека, который выразил мнение.

Само определение сентимент-анализа говорит об областях, в которых он находит применение, а определение мнения дополняет эту картину. Рассмотрим основные области.

Маркетинг. Сентимент-анализ позволяет компаниям получать статистику по продажам и отзывам пользователей о своем продукте. Эта информация может быть использована для индивидуализации продуктов или способов продвижения продуктов, либо для изменения самих продуктов. Например, пользователь в целом положительно отзывается о фотокамере, но отмечает недостаток в надежности. С помощью глубокого сентимент-анализа, компания может выделить отрицательную составляющую отзыва и поработать над надежностью продукта.

Политика. Мнения всегда составляли важнейшую часть политических влияний и систем. Можно сказать, политическая система базируется на основе системы мнений людей абсолютно разных классов и сфер. Сентимент-анализ, главным образом, используется в политике с целью анализа мнений народных масс с целью влияния.

Мониторинг мнений миллионов пользователей Сети. Развитие информационных технологий и, в частности, Интернета, позволяет упрощать процесс контроля и анализа народных мнений. Существует достаточное разнообразие различных инструментов, позволяющих автоматизированно извлекать необходимую информацию с различных порталов, предоставляя ее в удобном для анализа виде.

Развитие Интернета не только дает более простой доступ к мнениям пользователей, но и значительно увеличивает объем информации. Обрабатывать данные вручную – трудоемкий и ресурсозатратный процесс. Поэтому сентимент-анализ, в большинстве своем, проводится автоматизированно, с использованием следующих методов:

- методы, использующие правила либо шаблоны. Один из простейших видов анализа. Он заключается в применении к информации неких шаблонов и анализе полученных результатов. Сложность подобного анализа заключается лишь в составлении правил либо шаблонов, которые будут применяться в дальнейшем. Главным минусом подхода является его негибкость;
- методы машинного обучения (с учителем и без). Один из наиболее популярных видов сентимент-анализа. Эти методы более за-

тратны, чем шаблонные, поскольку требуют этапа обучения модели, однако и наиболее гибкие, ибо не привязаны к определенным языковым паттернам и доменам;

- методы, основанные на теоретико-графовых моделях. Суть этих методов заключается в предположении, что не все слова в документе равнозначно важны. Минусом методов является то, что слова, особенно в русском языке, зависимы от домена и могут иметь кардинально разные значения в разных доменах.

В своей работе я использовал три метода машинного обучения – наивный Байес (Bayes), метод опорных векторов (SVM) и рекуррентную нейронную сеть (RNN). Для RNN функцией активации использовалась ReLU (rectifier) – выпрямитель. Это наиболее простая функция, которая имеет следующий вид:

$$f(x) = \begin{cases} 0.01x, & x < 0 \\ x, & x \geq 0 \end{cases}$$

Ниже представлены полученные результаты:

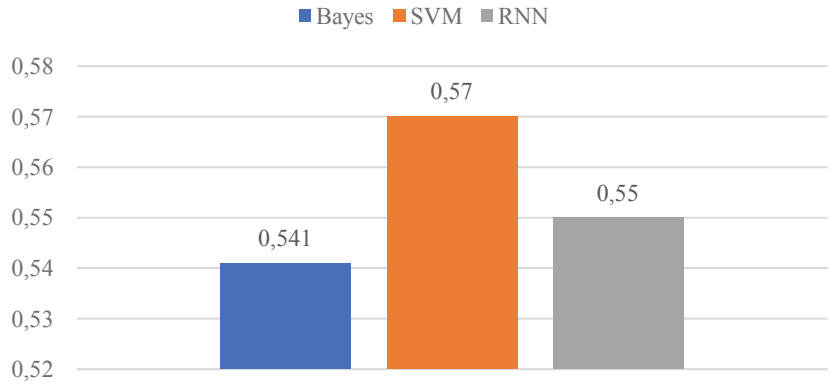

Рисунок 1 – Точность определения при 1000 записях

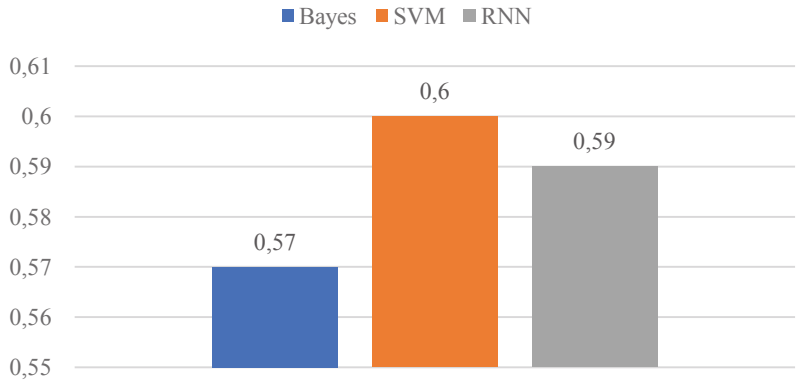

Рисунок 2 – Точность определения при 5000 записях

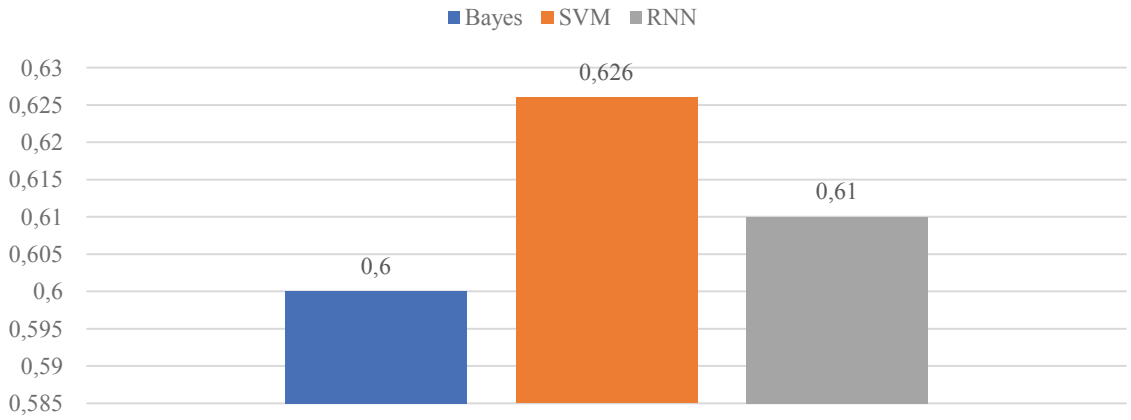

Рисунок 3 – Точность определения при 5000 записях с предобработкой

Для получения результатов на Рис.3 была проведена предобработка данных: удалены общие слова (меньше 4 букв), знаки пунктуации, использовались биграммы вместо униграмм.

По графикам можно сделать вывод, что лучший результат показывает SVM и RNN. Причем, качество RNN растет с увеличением обучающих данных. При достаточном количестве данных и увеличении классов с двух («положительный», «отрицательный») до трех («нейтральный») или переходе на оценочную систему, RNN выглядит лучшим вариантом.

Вместе с этим, можно выделить направления для улучшения результатов: комбинация n-грамм, применение TF-IDF, возможную комбинацию SVM и RNN.

ЛИТЕРАТУРА

1. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012. – 141 с.

2. Пазельская А., Соловьев А. Метод определения эмоций в текстах на русском языке, Москва, 2011. – 510-222 с.

3. Recurrent Neural Network Tutorial, Part 4 – Implementing GRU/LSTM RNN with Python and Theano [электронныйресурс] / WildML – Режимдоступа: http://www.wildml.com/2015/10/recurrent-neural-network-tutorial-part-4-implementing-a-grulstm-rnn-with-python-and-theano/ - Датадоступа 09.04.2019.