Группа №2
Мамонова Татьяна, Семенов Кирилл, Смирнова Александра, Торубаров Иван
Статья: On the Impact of SeedWords on Sentiment Polarity Lexicon Induction, ссылка: https://www.aclweb.org/anthology/C16-1147.pdf

Тестсет

все готовые файлы находятся здесь:
https://www.dropbox.com/sh/6imiguuzugv9qx8/AAAAZNFA4BM2N5Weq8PPcFP0a?dl=0

Реферат

Задача, которую решали авторы — доказать, что для sentiment lexicon может быть достаточно небольшое количество слов — сотни слов и фраз в качестве seed words. Такой подход дает два преимущества: 1) использование высококачественного словаря в качестве seeds, 2) их автоматический анализ с помощью затравки (bootstrapping). Авторами был проведен эксперимент по созданию sentiment lexicon для македонского языка. Стоит отметить со стороны еще одно преимущество, что раньше подобных исследований для македонского языка не проводилось.

Авторами используется подход составления sentiment-словаря, и они, говоря об иных опытах, перечисляют удачные по их мнению sentiment-словари, среди которых: LIWC (Pennebaker et al., 2001) — 2,300 слов; the General Inquirer (Stone et al., 1966) — 4,206 слов, Bing Liu's lexicon (Hu and Liu, 2004) — 6,786 слов, MPQA (Wilson et al., 2005) примерно 8000 слов. Крупномасштабные словари, такие как SentiWordNet (Esuli и Sebastiani, 2006; Baccianella et al., 2010); Hashtag Sentiment и Sentiment140, которые были разработаны командой NRC Canada в задании на SemEval-2013 по анализу настроений в Twitter. Авторы подчеркивают, что эти словари были составлены на большой базе "семян"; они хотят показать, что можно добиться не менее хороших результатов, используя "семена" среднего размера (исчисляя слова и фразы не в десятках тысяч, а в сотнях).

В качестве данных авторы использовали 500 000 постов в Твиттере на македонском языке. Тестовый датасет seed-words состоял из 1139 фраз и слов. Они были размечены двумя носителями македонского языка, которым требовалось определить, положительная, отрицательная или нейтральная оценка характерна для твита. В финальный тестовый датасет не вошли те данные, насчет которых у аннотаторов возникли разногласия, в итоге получилось 1088 фраз. Тренировочный датасет состоял из 8583 твитов, размеченных одним носителем македонского, он размечал характеристику каждого предложения внутри твита. Кроме этих датасетов были использованы sentiment-словари. Поскольку их не было в открытом доступе для македонского, авторы статьи взяли за основу Bing Liu's lexicon (2,006 положительных и 4,783 отрицательных слов), MPQA (2,718 положительных и 4,912 отрицательных слов), и болгарский словарь, состоявший из рецензий на фильмы, перевели их на македонский и скорректировали перевод.

Авторы статьи использовали несколько sentiment-словарей разного объема, чтобы понять, какой объем достаточен для sentiment-анализа. Таблица ниже их иллюстрирует.

Type of seed	Seeds	Unigrams	Bigrams	Total
Smileys: NRC	5	128	2,163	2,291
Words: Turney	10	865	14,343	15,208
Words: NRC	60	1,669	32,459	34128
Words: MCL	100	1,926	40,242	42,168
Words: MCL	200	3,752	60711	64,463
Words: MCL	500	7,219	124,977	132,196
Words: MCL	1,088	9,746	160,526	170,272

Table 2: Statistics about the lexicons we built using bootstrapping with PMI and LR. MCL is the manually-crafted lexicon.

Для обучения были важны следующие признаки: количество позитивных и негативных токенов, соотношение позитивных или негативных токенов к общему количеству токенов, сумма всех позитивных scores, всех негативных и общая сумма. Для классификации была использована логистическая регрессия, где основными признаками были значения TF.IDF для униграм, биграм и смайликов (или эмодзи). Кроме того, были различные дополнительные признаки, основанные на том, как часто в одном и том же твите появлялись позитивная и негативная лексика, на основе которых score пересчитывался.
Авторы статьи сравнивали то, как объем корпуса, его происхождение и метод увеличения объема (регрессия или PMI) влияют на результаты.

Работа авторов разделена на две подзадачи: создание расширенного (bootstrapped) лексикона на основе “семян” и обучение на основе этих лексиконов классификационной модели. Соответственно, авторы приводят результаты для каждой из подзадач.
В таблице, которую мы привели ранее, приведены результаты генерации лексиконов двумя способами из наборов “семян” разного объёма. Показано, что больший набор исходных слов приводит к более полному результирующему лексикону. При этом, хотя рост объёма лексикона и замедляется с ростом количества “семян” (если для минимального объёма он составляет 25x для униграмм и 433x для биграмм, то для максимального -- всего 9,5x и 148x соответственно), это замедление недостаточно значительное для того, чтобы можно было признать оптимальным любой из меньших исследованных объёмов “семян”.
Эта же таблица иллюстрирует значительную разницу в продуктивности двух методов генерации лексикона: униграммного (PMI) и биграммного (LR). Показано, что биграммный метод определяет на порядок больше таргетов: их вклад в работу моделей, впрочем, лучше демонстрируется в следующей таблице, описывающей результаты работы классификационных моделей.

Seeds for bootstrapping	Source	PMI			LR			PMI + LR
		B	B+S	B+S+M	B	B+S	B+S+M	B+S+M
–	–	–	61.99	78.18	–	61.99	78.18	78.18
1+1 words	(Turney, 2002)	51.48	62.29	78.40	59.82	63.57	78.51	78.89
2+3 smileys	(Mohammad et al., 2013)	61.12	63.81	78.69	65.18	68.99	78.95	79.62
5+5 words	manually-selected	62.55	64.59	79.25	66.70	69.73	80.13	80.87
7+7 words	(Turney and Littman, 2003)	63.02	66.27	79.71	66.98	69.82	80.54	81.99
30+30 words	(Mohammad et al., 2013)	63.47	68.51	79.84	67.28	70.01	80.68	81.33
50+50 words	our MCL	67.11	72.48	80.89	69.79	74.15	81.96	82.73
100+100 words	our MCL	70.94	76.30	82.76	71.41	77.47	84.72	85.45
250+250 words	our MCL	72.25	84.72	92.23	73.76	85.89	93.47	93.55
459+629 words	our MCL	73.82	90.91	94.12	75.29	91.02	94.32	94.44

Table 3: Sentiment classification results (F-score) using lexicons bootstrapped with PMI, LR, or both to calculate $SO(w)$: B = using the bootstrapped lexicon only, B+S = also using non-lexicon features, B+S+M = also using our MCL. The first line shows results when no bootstrapped lexicon is used.

Для оценки качества моделей авторы используют среднее между двумя F1-мерами, для которых целевыми переменными считаются соответственно тексты с позитивной и и негативной окраской. В строках таблицы представлены лексиконы, построенные на наборах “семян” различного объёма; в столбцах -- различные методы генерации лексиконов и обучения модели. Видно, во-первых, что униграммная и биграммная генерация лексикона дают приблизительно одинаковое качество, хотя биграммная в среднем всегда показывает метрику на 1% выше. Объединение униграммного и биграммного лексиконов также стабильно работает лучше, чем каждый лексикон в отдельности, хотя по сравнению с LR-лексиконом прирост ещё меньше (0,5%).
Значительно большую разницу между метриками вносит разница в объёмах наборов "семян" и использование в моделях дополнительных метрик (feature engineering), а также добавление в автоматически сгенерированый лексикон исходных "семян".
Скомбинировав все доработки модели на самом большом лексиконе, авторы получили наилучший результат со значением метрики 94,44.
Несмотря на довольно обширное обсуждение результатов, авторы не рассматривают проблемы алгоритма и не приводят примеров неверной работы моделей.

Безусловно, большой плюс этой статьи в том, что описывается не только технология увеличения лексикона при помощи seed words, но и весь пайплайн обработки текстов для дальнейшего сентимент-анализа. Впрочем, увы, эти пайплайны очень зависят от языка, и для русского большое количество рассмотренных эвристик (вроде стемминга) не подойдет.
Очень корректно подчеркивалось, что задача бинарной классификации sentiment-orientation для каждого слова является на самом деле классификацией на три группы - положительные, отрицательные и нейтральные.
Что касается особенностей данного проекта, то достаточно неплохой эвристикой можно считать использование sentiment lexicons из других языков. В данном случае для составления македонского словаря использовались sentiment lexicons английского, русского и болгарского языков. Безусловно, такой шаг стоит делать с осторожностью, потому что в различных языках у слов с прямым переводом есть различные коннотации; но получить таким образом несколько сотен полезных слов вполне возможно. Еще более важен их вывод: подготовленные таким образом

лексиконы всегда дают худший результат, чем вручную выбранные для данного языка слова.
Минус - нет псевдокодов. Без этого было сложно понимать, как взаимодействуют алгоритмы PMI и Logistic Regression при обучении, и что происходит со словами, которые образуют пересечение множеств иностранных переводов и вручную придуманных лексиконов. Псевдокод очень помог бы в п. 4.3 и п. 6.
Не очень понятным осталось три момента. Во-первых, мы так и не смогли понять, что значит mid-sized seed, учитывая, что самые хорошие результаты получались у исследователей при количестве seedwords чуть больше тысячи, а во введении описано, что лексиконы в принципе делают обычно объемом меньше, чем 10 000. Можно ли в итоге их выборку считать mid-sized - вопрос открытый. Во-вторых, при подведении результатов в таблицах и диаграммах стало не очень ясно, в чем оппозиция между B, S и M: например, какие взаимные пропорции между B и M (когда предлагается сложить B+S+M). В-третьих, не очень понятно, что именно входило в дополнительные признаки при построении модели логистической регрессии.