Технічні науки

УДК 004.852

Олашин Олександр Олександрович
студент
Національний технічний університет України
«Київський Політехнічний Інститут»
Шипік Данил Володимирович
студент
Національний технічний університет України
«Київський Політехнічний Інститут»
Олашин Александр Александрович
студент
Национальный технический университет Украины
«Киевский политехнический институт»
Шипик Данил Владимирович
студент
Национальный технический университет Украины
«Киевский политехнический институт»
Olashyn Oleksandr
student
National Technical University of Ukraine
«Kyiv Polytechnic Institute»
Shypik Danil
student
National Technical University of Ukraine
«Kyiv Polytechnic Institute»

ПОРІВНЯННЯ ТОЧНОСТІ АЛГОРИТМІВ АНАЛІЗУ ТОНАЛЬНОСТІ НА ПРИКЛАДІ ТВІТТІВ

СРАВНЕНИЕ ТОЧНОСТИ АЛГОРИТМОВ АНАЛИЗА ТОНАЛЬНОСТИ НА ПРИМЕРЕ ТВИТТОВ

COMPARISON OF ACCURACY OF SENTIMENT ANALYSIS ALGORITHM ON TWITTER MESSAGES

Анотація: Проведено порівняння ефективності роботи різних алгоритмів аналізу тональності (а саме наївного баєсового класифікатора та методу опорних векторів) на виборці коментарів з твіттеру. Зроблено висновки щодо ефективності їх застосування.

Ключові слова: аналіз тональності, наївний баєсів класифікатор, твіттер, метод опорних векторів, біграми.

Аннотация: Проведено сравнение эффективности работы разных алгоритмов анализа тональности (а именно наивного байесовского классификатора и метода опорных векторов) на выборке комментариев из твиттера. Сделано выводы об эффективности их применения.

Ключевые слова: анализ тональности, наивный байесовский классификатор, твиттер, метод опорных векторов, биграммы.

Summary: A comparison of the efficiency of different sentiment analysis algorithms (naive bayes classifier and SVM) was made using a set of twitter comments. A general conclusion is made concerning the effectiveness of these algorithms.

Key words: sentiment analysis, naive bayes classifier, twitter, SVM, bigrams.

Аналіз тональності тексту (англ. Sentiment analysis) є відносно новим напрямком автоматизації аналізу емоційної складової тексту. Він набуває популярність у зв’язку з розвитком різних платформ для оцінювання (будь-то сайт про фільми, одяг чи техніку). Правильне його застосування дозволяє оцінити реакцію користувачів на той чи інший продукт і врахувати її в подальшому [1, с. 79; 2, с. 2545-2546].

Однак проблемою такого аналізу є те, що не завжди можна просто визначити точне емоціне забарвлення тексту опираючись тільки на окреме слово. Поширене використання набули емотікони та абревіатурні скорочення, які в сукупності можуть нести зовсім інший емоційний зміст ніж по одинці. Або ж текст може містити велику кількість негативних або позитивних слів і все одно виражати зовсім протилежну думку [2, с. 2544-2545, с. 2547-2548].

Тому одним з напрямків аналізу тональності тексту є вибір методів таким чином, щоб проводити класифікацію максимально точно, врахавуючи різні можливі комбінації [2, с. 2546].

В нашій роботі ми зосередили свою увагу на двох розповсюджених алгоритмах: наївному баєсовському класифікаторі та методі опорних векторів. Перший використовує теорему Баєса для визначення ймовірності приналежності елементу спостереження до одного з наперед заданих класів. Недоліком цього методу (через який він і називається «наївним») є те, що ми вважаємо, що слова зустрічаються незалежно, що в загальному випадку не є вірним. Однак в реальних умовах він є досить ефективним, і має досить багато плюсів - швидкодія, простота, помірні вимоги до пам’яті, через що він набув доволі широкого розповсюдження [4, с. 1,6].

Метод опорних векторів (англ. SVM – тут і надалі буде застосовано це скорочення через розповсюдження в літературі) - це метод класифікації, що визначає класи за допомогою меж просторів. Тобто вихідні вектори переводяться в простір більш високої розмірності і шукаються роздільні гіперплощини з максимальним проміжком між ними. Цей метод належить до розряду лінійних класифікаторів. Його перевагами є те, що SVM дозволяє отримати рішення близьке до оптимального, навіть без вбудованих знань про предметну область, при чому завдяки тому, що цей метод зводиться до вирішення задачі квадратичного програмування на випуклому просторі - він гарантує єдиність розв'язку. Серед недоліків методу - значне збільшення обчислювальної складності при збільшенні ефективності [3, с. 417-429, 434-436, 443-444].

Для покращення роботи алгоритмів застосовуються біграми (n-грами з n=2). Біграма - це послідовність з двох елементів (в нашому випадку слів). Вони враховуються в алгоритмі, як один змістовний елемент [6].

Для реалізації наведених вище алгоритмів було використано мову python та бібліотеки nltk та sci-kit learn. Робота програм проводилась на вибірці коментарів з твіттеру розміром близько 1.5 млн коментарів. Всі вони були підготовлені для використання (тобто містили емоційну оцінку - позитивну чи негативну) [5]. Біграми були знайдені за допомогою BigramCollocationFinder з бібліотеки nltk з параметрами - ширина ковзного вікна - 4, фільтр частоти (мінімальна кількість кандидата на біграми в тексті) - 3, а кількість обраних біграм - 1000.

Отримані результати демонструє наступна таблиця:

Таблиця 1

Порівняння отриманих характеристик використаних методів

	Precision	Recall	Accuracy	Фальшиво позитивні	Фальшиво негативні
НБК	0,6805	0,8717	0,7650	0,0750	0,1599
SVM	0,5720	0,8253	0,7829	0,0928	0,1241
НБК* з біграмами	0,6852	0,8740	0,7649	0,0739	0,1610
SVM з біграмами	0,5256	0,7973	0,7813	0,1037	0,1149

де НБК - наївний баєсовьский класифікатор,

$$Precision = \left(\frac{TP}{TP+FN}\right); Recall = \left(\frac{TP}{TP+FP}\right); Accuracy = \left(\frac{TP+TN}{P+N}\right)$$

TN (TP) - кількість дійсно негативних (позитивних) коментарів; FN (FP) - кількість фальшиво негативних (позитивних) коментарів (тобто коментарів, що були невірно віднесені до певного класу); N (P) - кількість негативних (позитивних) коментарів. Отримані показники описують: precision - наскільки точним був результат пошуку; recall - наскільки повним був результат; accuracy - доля правильних відповідей [7, с. 39].

Висновки. Результати приведені в таблиці 1 свідчать про те, що: враховуючи precision, recall та швидкодію, що істотно більше для баєсовського алгоритму ніж для SVM (precision - на 11%, recall - 6%, а для випадку з використанням біграм - 5% та 8% відповідно) і незважаючи на програш в accuracy (приблизно 2% в обох випадках) на думку авторів, наївний байєсовський класифікатор виявився кращим. Можна також сказати,

що будь який з цих методів показує результат, що значно кращий за випадковий вибір - 50% (оскільки у виборці однакова кількість двох класів)

Також, отримані дані свідчать про те, що додавання біграм в випадку наївного байесовського класифікатора не дуже сильно покращує результат (збільшення precision - на 0,5%, recall - на 0,2% і зменшення accuracy - на 0,1%), а для SVM навіть його погіршує (зменшення precision - на 5%, recall - на 3% і accuracy - на 0,15%). Отримані результати відносно наївного баєсовського класифікатора досить гарно корелюють з [1, с. 85].

Література:

1. Bo Pang, Lillian Lee, Shivakumar Vaithyanathan Thumbs up? Sentiment Classification using Machine Learning Techniques / Bo Pang, Lillian Lee, Shivakumar Vaithyanathan // Proceedings of EMNLP.-2002.-C. 79-86

2. Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, Arvid Kappas Sentiment strength detection in short informal text / Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, Arvid Kappas // Journal of the American Society for Information Science and Technology №61.-2010-C. 2544–2558

3. Хайнакин С. Нейронные сети. Полный курс. Второе издание. - М: Издательский дом «Вильямс», 2006.- 1104с.

4. Irina Rish An empirical study of the naive Bayes classifier (2001) [Електронний ресурс] / Irina Rish - Режим доступу: http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf

5. Twitter Sentiment Analysis Training Corpus (Dataset) [Електронний ресурс] / thinknook.com - Режим доступу: http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/

6. Kavita Ganesan What are N-Grams? [Електронний ресурс] / Kavita Ganesan - Режим доступу: http://www.text-analytics101.com/2014/11/what-are-n-grams.html

7. David M W Powers Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation / David M W Powers // Journal of Machine Learning Technologies №2 (1).-2011-С.37–63