Застосування наївного байесовського класифікатору для визначення емоційного змісту текстових даних

А.Б.Бєгунов, к.т.н. Т.М.Заболотня

Національний технічний університет України “КПІ”
Україна, Київ
arxton@mail.ru

Оцінювання тональності, або емоційного змісту, текстових даних є одним з важливих аспектів аналізу природномовних повідомлень у сучасних інформаційних системах. Під тональністю тексту зазвичай розуміють позитивне, негативне або нейтральне забарвлення як цілого текстового документу, так і його окремих частин, які мають відношення до певних понять, таких як персони, організації, бренди тощо [1]. Існуюче програмне забезпечення реалізує автоматизований аналіз емоційного змісту тексту виключно за ступенем позитивності, що робить результати роботи таких програм досить грубими і значною мірою звужує коло сфер їх застосування. Наприклад, якщо текстове повідомлення негативне, то воно може виражати страх або агресію, що не є тотожними емоціями. Виправити дану ситуацію можна шляхом розширення спектра тональностей, які здатна розрізняти програмна система. Таким чином, налаштування програмного забезпечення на визначення багатовекторної картини емоційного забарвлення тексту є цікавою та актуальною задачею.

У доповіді запропоновано метод визначення емоційного змісту текстових повідомлень на основі наївної байесовської моделі. Згідно методу, текстові дані складаються з набору термів (слів) t_1, $t_2,\ldots, t_n$. Емоційний зміст тексту презентується набором категорій (класів) емоцій (наприклад, страх, задоволення тощо): $C_1, C_2,\ldots, C_m$. Для наївної байесовської моделі робиться припущення про статистичну незалежність характеристик документів (слів) $t_1, t_2,\ldots, t_n$. Для кожної категорії C_j на вхід системи визначення емоційного змісту на етапі навчання подається еталонний документ E_j, що містить текстові дані, притаманні цій категорії. Далі на основі цих документів визначаються коефіцієнти, що характеризують умовну ймовірність приналежності слова t_i до відповідної категорії. Нехай треба проаналізувати документ D, що характеризується словами $t_1, t_2,\ldots, t_n$. Позначимо $p_1^j, p_2^j,\ldots, p_n^j$ як відповідні коефіцієнти для слів з документу D відносно класу C_j. Тоді за формулою $S^j = p_1^j \cdot p_2^j \ldots p_n^j / (p_1^j \cdot p_2^j \ldots p_n^j + (n-1)(1-p_1^j)\cdot(1-p_2^j) \ldots (1-p_n^j))$ можна розрахувати ступінь S^j входження документу D до класу C_j. Входження даних до будь-якого класу C_j тут приймається однаково вірогідним.

Для реалізації вищеописаного методу визначення емоційного змісту текстових даних створено програмне забезпечення автоматизованого оцінювання тональності тексту для кожної з категорій емоційного забарвлення. В якості результату дана програма повертає вектор характеристик $S^1, S^2,\ldots, S^m$, які відображають ступінь входження документу до кожного з класів $C^1, C^2,\ldots, C^m$. Перевагою цієї системи є гнучкість і розширюваність: гнучкість полягає в можливості впливу на коефіцієнти системи за рахунок зміни еталонних документів, а розширюваність - в можливості введення будь-якої кількості нових категорій емоцій до аналізу в процесі функціонування системи. У доповіді викладено основні алгоритми функціонування даної розробки, а також наведені результати її застосування до тестового масиву текстових повідомлень.

Таким чином, реалізація запропонованого методу сприяє підвищенню точності роботи програмного забезпечення автоматизованого визначення емоційного змісту текстових даних і може бути використана при вирішенні широкого спектру задач, зокрема, для комп’ютеризованого аналізу впливу інформації із ЗМІ на людей, аналізу психоемоційного стану колективу у крупних корпораціях тощо.

ВИКОРИСТАНІ ДЖЕРЕЛА

1. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы.- М.: Либроком, 2009.-264с.
2. David D. Lewis Naive (Bayes) at forty: the independence assumption in information retrieval. In Proceedings of the ECML-98, Chemnitz, DE: Springer Verlag, Heidelberg, DE. – 1998. – P. 4–15.