ПОДОБРЯВАНЕ НА СПИСЪКА ОТ ПОЛОЖИТЕЛНИ И ОТРИЦАТЕЛНИ ТЕРМИНИ ПРИ ИЗВЛИЧАНЕТО НА МНЕНИЯ ОТ ТЕКСТ

Тодор Цонков[1], Иван Койчев[1,2]

[1]Факултет по математика и информатика,
Софийски университет „Св. Климент. Охридски”
[2]Институт по математика и информатика - БАН
todort@fmi.uni-sofia.bg, koychev@fmi.uni-sofia.bg

***Резюме**: Целта на представената разработка е да предложи алгоритъм за подобрение на списъка от думи и изрази носещи определено отношение (положително или отрицателно) към даден обект. Този списък се използва по-късно за автоматична класификация на нови мнения за обекти. Алгоритъма използва мнения класифицирани от човек след това ги анализира за да намери най-често срещаните думи и изрази, които не са в списъка от “спиращи думи”. Такъв списък се изграждат за определена предметна област. Проведените експерименти с текстове от различни предметни области, показват че предложният подход подобрява точността на класификацията на мнения.*

***Ключови думи**: извличане на мнения от текст, списък от положителни и отрицателни думи*

1. Увод

Задачата за определяне на чувства в текст е обект на сериозен изследователски интерес през последните години. Предпоставка за това са все по-голямото използване на социалните мрежи (Facebook, Twitter и др.), блогове и форуми. Това налага създаването и използването на методи, които да оценяват дали едно мнение е положително, отрицателно или неутрално. Тези методи имат различни приложения в маркетинг анализа, мнението на хората относно политиката и т.н. Например, ако искаме да закупим лаптоп освен цена, характеристики, производители т.н. биха ни интересували мненията на други хора (най-вече такива, които считаме за експерти в областта), които биха ни дали съвети. За производителите на такива лаптопи мнения на клиентите са изключително полезни за уточняване на маркетинг стратегията и за улавяне на погрешни мнения - например “На този лаптоп не върви новият Уиндоус.” [1][2]

Въпреки че на пръв поглед задачата изглежда лесна изречения от рода на “Този филм би трябвало много да ми хареса, актьорите са невероятни, сюжета изглежда повече от обещаващ. Но накрая нещо не се получава.” са много

трудни за класификация от машина, тъй като се съдържат предимно думи, носещи положително значение.

Една от основните стъпки при извличането на чувства от текст е използването на ключови думи, които да носят позитивен, негативен или неутрален контекст. Например думи като “харесвам”, “обичам”, “мразя” или “не харесвам”. На базата на списък от тези думи може чрез различни алгоритми (например наивен бейсов класификатор) да определим чувството в текста.

Възможно е да се използват различни класификатори – например support vector machines, наивен бейсов и други. [3]

Идеята на настоящата статия е да предложи алгоритъм за подобрението на този списък, тъй като освен думите, които носят положително или отрицателно отношение във всяка предметна област (като харесвам, мразя и т.н.) се срещат и други думи или поредица от две думи, които да носят определено отношение в зависимост от конкретната предметна област (например в областта на футбола - “отборът се защитаваше много” носи отрицателно значение, а никоя от думите сама по себе си не е с такова). Характерна особеност за социалните мрежи (например Twitter) е, че някои съкращения, например тагове, могат да носят сами по себе си значение за мнението в текста.

В настоящата статия са направени тестове в няколко предметни области – футбол и политика като използваме социалните мрежи Facebook, Twitter, Google+, Blogspot, Digg, Yahoo News. Използвани са стоп думи от https://code.google.com/p/twitter-sentiment-analysis/source/browse/trunk/files/stopwords.txt?r=51.

Разгледано е определяне на резултати с помощта на класификатора SentiStrength (който може да се използва за научни цели) в/у 1000 произволни мнения от социалните мрежи. Показва се, че точността на класификация се подобрява с няколко процента при ползването на подобрения речник с ключови думи.

Класификацията на мненията на позитивни/негативни/неутрални е направена от авторите на статията с помощта на програмни средства за верификация на получените резултати (SentiStrength) и верификация на резултатите от авторите. т.е. Всяко мнение класифицирано като положително и отрицателно е проверено от авторите.

2. Подобни разработки

През последните години се наблюдава засилен интерес в областта на извличането на чувства от текст. Ранната работа в тази насока е на Turney и Pang , които прилагат различни методи за намиране на полярността на ревюта на продукти и на ревюта на филми. Това се прилага на ниво документ.

Документите могат да бъдат класифицирани и на степени на отношение - например в скалата от (-5; +5) изречението:

“Много харесвам този продукт - най-добрият, който някога съм ползвал!”

може да се оцени като +5, а мнение от рода на “Това не е лошо” да се оцени като +1.

Изречение от рода на “Изключително много мразя този продукт!!!” може да бъде оценено като -5. В настоящата разработка се спираме изцяло на класификация - положително/отрицателно/неутрално (т.е. дали е засечено отношение или не). Pang и Snyder разглеждат точно такъв тип класификация.

За верификация на получените резултати е използван софтуерният инструмент SentiStrength (http://sentistrength.wlv.ac.uk/), който има резултати, достигащи класификационните способности на човек. Доказано е, че човек има средно 79% успеваемост в класифицирането на мнения и дори програма да има 100% точност, двама произволни човека ще имат разлика в класификацията около 20% на произволно избрани мнения (както и на която и да е тема). Горепосоченото софтуерно приложение осигурява Java код, който е базиран на списъци от думи, носещи положително и отрицателно значение. В началото се използват списъците с позитивни/негативни думи, които предлага SentiStrength, които се обогатяват с всяка предметна област. В настоящият алгоритъм не се засичат т.нар. emotion icons, което би могло да бъде подобрения, тъй като в доста мнения (като “I’m walking outside :)))))))))))))))))”) отношение може да бъде засечен. Редица софтуерни продукти с отворен код използват решения като машинно самообучение, статистка, и обработка на естествен език, за да автоматизират извличането на чувства от текст в уеб страници, онлайн новини, дискусионни групи в Интернет, онлайн ревюта, уеб блогове и социалните медии. Системи, базирани на знания ползват публично достъпни ресурси като WordNet-Affect, SentiWordNet и SenticNet, за да извличат семантичната информация с методи на обработката на естествен език.

3. Алгоритъм за подобряване на лексикона от думи

Списък позитивни думи = {Предварително зададен списък от думи, които носят универсално положително значение}

Докато списъкът от думи не се променя повтаряй:

1. Извлечи произволни 1000 положителни кратки изречения от социалните мрежи (примерно от определена област или по ключова дума). Класификацията се извършва от учител.
2. Намери най-често срещаните N думи в изреченията, класифицирани като позитивни, които не се съдържат в даден списък от стоп думи.

3. Обнови списъка с думи като добавиш тези, които се срещат по-често от досегашните на базата на отношение - срещане/брой думи в изреченията.

Алгоритъмът трябва да бъде направен и за отрицателни думи по същия начин. Той е изцяло еквивалентен на гореописания. Следва илюстрация на алгоритъма и за думите с отрицателно отношение:

1. Добави най-често срещаните думи с положително отношение.
2. Добави най-често срещаните думи с отрицателно отношение.
3. Класифицирай отново с помощта на някой класификатор, ползващ списък от думи с положително и отрицателно значение и сравни резултатите с вече получените.

Предметната област на тестовете може да бъде зададена предварително. Алгоритъмът разчита на потребителя да бъде обучен от предварително зададена тема.

Да дадем следния пример – Извличаме по 1000 мнения от социалните мрежи, които съдържат дадена дума (“football” или “politics”), при което намираме следните подобрения в списъка - думи като “scored”, “goal” носят положително значение, докато думи като “defended” или дори биграма като “zonal marking” може да носят по-скоро отрицателно отношение. От всички извлечени мнения броим най-често срещаните думи и биграми, които не попадат в списъка от стоп думи и ги добавяме към списъка от думи. След това тестваме дали има подобрение на класификацията отново с 1000 произволно извлечени мнения от предметната област. Предметната област може да бъде лесно заменена, посочените примери са в области, които са с широк публичен интерес, всички мнения се извличат на база ключова дума. Може да се направи, така че да се търсят биграми или поредица от думи, но тогава те трябва да бъдат добавени в списъка от получените стоп думи, тъй като се срещат във всяко мнение и не носят значение.

4. Експерименти

Гореописаният алгоритъм в секция 3 е тестван в следните предметни области - футбол и политика. След изпълняването на алгоритъма и подобряване на списъка от думи в областта на футбола се подобрява класификацията с наивен бейсов класификатор със средно 3,5%, а в областта на политиката с около 8%. Посочената разлика се дължи на това, че в политиката повече думи носят отрицателно значение.

Използвани са социалните мрежи Facebook, Twitter, Google+, Digg, Bing Search като се извличат най-актуалните мнения при всяко тестване. Почти всички от социалните мрежи предлагат възможност за извличане на публични

мнения от тях като лесно могат да се добавят социални мрежи, тъй като в повечето случаи извличаният формат е един и същ - JSON.

При всеки тест извличаме произволни 1000 мнения от социалните мрежи. Разпределението по социални мрежи е следното:

Приблизително 600 (т.е. 60%) от мненията са от Twitter. Приблизително 300 (т.е. 30%) са от Фейсбук. Останалите 10% са от другите мрежи в зависимост от най-новите резултати. Всички мнения са сортирани по дата (т.е. най-нови) и са публични, т.е. имаме права да ги използваме. Използваме най-често Twitter и Facebook, тъй като има най-много мнения, публикувани там.

След анализ на футболни мнения открихме, че най-често срещаните думи, които са специфични за областта и не са стоп думи: ball, cup, player, Match, Win, Lose, Play, Team, Goalkeeper, Striker, Goal, Kick, Pass, Tackle, Cross, Dribble, Shoot, Strike,Score, Foul, Defend, Attack, Referee, Penalty, Red.

От изброените по-горе думи положително значение в контекста на областта носят:cup,goal,strike,score,team,shoot.

Думите goalkeeper, lose, referee, red, penalty носят отрицателно значение.

След добавянето им в списък с думи с общо положително значение при класификацията с помощта на софтуерният инструмент SentiStrength се постига подобрение от 4% в получените резултати - от 71% на 75% и от 66% на 71%.

В областта на политиката голяма част от мненията се класифицират като негативни (около 40%) като средното за другите области е около 10-15%. С добавянето на специфични термини се увеличава точността на предсказване с около 8% - например най-често срещаните думи и биграми са: politician, tax, elections, president, minister, new president, tax cut.

5. Заключение

В настоящата разработка се изследва метод за подобряване на резултатите при извличането на чувства от мнения, базиран на извличане на специфични думи от предметна област и добавянето им. С направените тестове се вижда, че се подобряват получените резултати. За бъдещо надграждане ще бъдат разгледани други предметни области, ще се подобри избирането на специфичните думи и ще се тества с по-голямо множество от мнения.

Литература

1. Kraychev, B. and Koychev, I. Computationally Effective Algorithm for Information Extraction and Online Review Mining. In Proc. of International Conference on Web Intelligence, Mining and Semantics June 13-15, 2012, Craiova, Romania, ACM.

2. Kraychev, B. and Koychev, I. Classification of Online Reviews by Computational Semantic Lexicons. Proc. of Int. Conference S3T'11 (Track Intelligent Content and Semantic), Burgas 1-3 September 2011, Advances in Intelligent and Soft Computing series, Springer.

3. Pang, Bo; Lee, Lillian (2008). "4.1.2 Subjectivity Detection and Opinion Identification". *Opinion Mining and Sentiment Analysis*. Now PublishersInc.

IMPROVING THE LEXICON OF POSITIVE/NEGATIVE WORDS AND BIGRAMS FOR SENTIMENT ANALYSIS

Todor Tsonkov[1], Ivan Koychev[1,2]

[1]Sofia University "Kliment Ohridski"
[2]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
ttsonkov@gmail.com, koychev@fmi.uni-sofia.bg

Abstract: *The idea of the current paper is to propose an algorithm for improving the list of positive and negative words based on a specific topic. The opinions are classified by a person or a machine (or combined) and the most frequent words are being found that are not in the list of stop words. These words are being added to the list and than with a sample classifier is found improvement in the classification of the already exctracted opinions. Several tests have been described to show how to test the algorithm.*