Список літератури

1. Овчаров, С.В. Исследование потерь в асинхронном электродвигателе [Текст] / С.В. Овчаров, Р.В. Телюта // Вісник Харківського національного технологічного університету сільського господарства імені Петра Василенка. «Проблеми енергозабезпечення та енергозбереження в АПК України». – Харків: ХНТУСГ, 2009. – Вип. 86. – 5 с.
2. Исследование потерь активной мощности в асинхронном электродвигателе в функции скольжения [Текст] / В.В. Овчаров, С.В. Овчаров, Р.В. Телюта, О.В. Юдина // Вестник национального технического университета «ХПИ». – Харьков: НТУ ХПИ, 2009. – 8 с.
3. Телюта, Р. В. Методика экспериментального исследования потерь активной мощности в асинхронном электродвигателе / Р. В. Телюта // Пр. Тавр. держ. агротехнол. ун-ту. – 2011. – Вип. 11, т. 3. – С. 133-136.

Ruslan Teliuta, Assoc. Prof., Phd tech. sci., Serhei Plieshkov, Assoc. Prof., Phd tech. sci.
Central Ukrainian National Technical University, Kropivnitsky, Ukraine

The principles ofresearch model constructionfor active-power losses is in an asynchronous electric

The principles ofresearch model constructionfor active-power losses in an asynchronous electric motor is shown in the article.

In process, the offered methodology of research of active-power losses is in an asynchronous electric motor with a short-circuited rotor in the function of active-power, that were transmitted to from the billow of electric motor the working machine. Losses of active-power in an electric motor calculated as a difference of active-power, which consumed by an electric motor and transmitted power on the billow of electric motor. The first power measured by complete set of control and measuring devices, the second calculated in analytical way by means of parameters of G - similar substituting chart for one phase of asynchronous electric motor taking into account the height of pure resistance of puttee at heating and measured sliding.

The offered methodology of research of active-power losses in an asynchronous electric motor allowed setting that deviations of experimental data from analytical calculations are within the limits of ±5,88 %.

chart of substitution, parameters of chart of substitution, sliding, active watts-in, coefficient of losses
Одержано 18.05.17

УДК 004.9

Д.В. Шингалов, асп., Є.В. Мелешко, доц., канд. техн. наук, Р.М. Минайленко, доц., канд. техн. наук, В.А. Резніченко, викл.
Центральноукраїнський національний технічний університет, м. Кропивницький, Україна
E-mail: elismeleshko@gmail.com

Методи автоматичного аналізу тональності контенту у соціальних мережах для виявлення інформаційно-психологічних впливів

У статті проводиться дослідження існуючих методів автоматичного аналізу настроїв користувачів в онлайнових віртуальних соціальних мережах, для виявлення найбільш придатних з них для розпізнавання інформаційно-психологічних впливів, заснованих на емотивній комунікації. Дослідження включає в себе огляд лексемного методу та методів машинного навчання. Зокрема розглянуто процедуру попередньої обробки тексту перед аналізом та категорії словників, що використовуються для аналізу. Наведено алгоритми навчання для класифікації тексту: метод опорних векторів, наївний класифікатор Бейєса, дерева прийняття рішень, метод максимальної ентропії та нейронні мережі. Запропоновано методи для поліпшення якості результатів автоматичного визначення тональності тексту.

соціальні мережі, аналіз настроїв, лексемний аналіз, машинне навчання

© Д.В. Шингалов, Є.В. Мелешко, Р.М. Минайленко, В.А. Резніченко, 2017

Д.В. Шингалов, асп., Е.В. Мелешко, доц., канд. техн. наук, Р.Н. Минайленко, доц., канд. техн. наук, В.А. Резниченко, преп.
Центральноукраинский национальный технический университет, г. Кропивницкий, Украина

Методы автоматического анализа настроений в социальных сетях

В статье проводится исследование существующих методов автоматического анализа настроений пользователей в онлайновых виртуальных социальных сетях, для выявления наиболее подходящих из них для распознавания информационно-психологических воздействий, основанных на эмотивной коммуникации. Исследование включает в себя обзор лексемного метода и методы машинного обучения. В частности рассмотрено процедуру предварительной обработки текста перед анализом и категории словарей, используемых для анализа. Приведены алгоритмы обучения для классификации текста: метод опорных векторов, наивный классификатор Байеса, деревья принятия решений, метод максимальной энтропии, и нейронные сети. Предложены методы для улучшения качества результатов автоматического определения тональности текста.

социальные сети, анализ настроений, лексемного анализ, машинное обучение

Актуальність. Зростаюча популярність онлайнових соціальних мереж, таких як Twitter, Facebook, MySpace тощо, в останні роки викликало великий інтерес до аналізу настроїв їх користувачів. Аналіз тональності (настроїв) відноситься до класу методів, заснованих на обчислювальній обробці, що використовується для ідентифікації, вилучення та характеристики суб'єктивної інформації, наприклад, думок, виражених в тій чи іншій частині тексту, що дозволяє класифікувати ставлення автора до різних тем в позитивні, негативні або нейтральні категорії. Аналіз настроїв має безліч застосувань в різних галузях, включаючи бізнес-аналітику, політику, соціологію і т.д. Такі дані, як коментарі, блоги, твіти, відео і т.д., які висловлюють думки відносно різних тем і подій, пропонують величезні можливості для вивчення і аналізу людських думок і настроїв.

Постановка завдання. Метою даної роботи є дослідження існуючих методів автоматичного аналізу настроїв та тональності мультимедійних даних і текстів у віртуальних соціальних мережах для виявлення найбільш придатних з них для розпізнавання пропаганди, агітації та інформаційно-психологічних впливів у контенті соціальних мереж, заснованих на емотивній комунікації.

Задача дослідження. На основі дослідження параметрів та характеристик методів автоматичного аналізу тональності контенту віртуальних соціальних мереж, таких як необхідність застосування словників, необхідність попередньої лінгвістичної обробки тексту та можливість застосування до різних типів даних, виділити найбільш придатні методи для розпізнавання пропаганди та інформаційно-психологічних впливів у контенті.

Зв'язок проблеми з важливими науковими та практичними задачами. Розглянута проблема пов'язана з важливими науковими та практичними задачами лінгвістичного аналізу тексту, аналізу соціальних мереж та машинного навчання. Вирішення проблеми якісного автоматичного аналізу настроїв у соціальних мережах дозволить автоматично виявляти інформаційно-емоційні впливи у контенті соціальних спільнот та вчасно реагувати на дані впливи.

Виклад основного матеріалу. Сучасна теорія аналізу та керування великими даними (Big Data) виокремлює два основних напрямки автоматичного аналізу настроїв – це методи на основі використання лексем і методи машинного навчання.

Перед застосуванням будь-якого з методів вилучення настрою, звичайною є практика попередньої обробки даних. Попередньо оброблені дані дозволяють забезпечити високу якість класифікації тексту і зменшити обчислювальну складність. Типова процедура попередньої обробки включає в себе наступні основні кроки:

– Розмітка за частинами мови. Цей процес дозволяє автоматично визначити кожне слово речення як частину мови: іменник, займенник, прислівник, прикметник,

дієслово, вигук і т.д. Мета полягає в тому, щоб витягти зразки тексту на основі аналізу частотних розподілів цих частин у мові.

– Зведення до кореня. Процедура відсікання суфіксів та закінчень від кореня. Кількість різних слів для аналізу зменшується, коли корінь схожих слів, наприклад, таких як "читати", "читач" і "читання" відображаються як одне слово “читати”.

– Видалення некорисних слів. Це слова, які несуть в собі сполучну функцію в реченнях, наприклад, прийменники, артиклі і т.д. Немає певного списку таких слів, але деякі пошукові машини, не використовують такі слова як, "є", "в", "який" і "на". Ці слова можуть бути видалені з тексту перед класифікацією, так як вони мають високу частоту появи в тексті, але не впливають на його емоційне навантаження.

– Обробка заперечень. Заперечення відноситься до процесу перетворення настроїв тексту з позитивного на негативний, або з негативного на позитивний.

– Токенізація в N-грами. Токенізація – це процес створення словнику зі слів тексту. Токенізація соціальних медіа-даних значно складніша, ніж токенізація звичайного тексту, так як вони містять численні смайли, URL-посилання, скорочення, які не можуть бути легко оцінені в тексті.

Лексемно-орієнтований підхід обчислює настрій заданого тексту в залежності від полярності слів або фраз у цьому тексті. Методика розрахунку настрою [1] полягає у наступному: після попередньої обробки тексту, відбувається перевірка маркера кожного слова на його полярність в лексиконі. Якщо слово не знайдено у лексиконі, тоді його полярність вважається нульовою. Після призначення балів полярності W всім словам, що містяться у тексті, остаточна оцінка S настрою тексту розраховується (1) діленням суми балів слів, які задають настрій тексту (крім нульових) на кількість m таких слів:

$$S = \frac{1}{m}\sum_{i=1}^{m} W_i , \qquad (1)$$

де W_i – бал полярності i-го слова; m – кількість слів, які задають настрій тексту.

Усереднення балу дозволяє отримати числове значення балу настрою у діапазоні від -1 до 1, де 1 означає сильний позитивний настрій, -1 означає сильний негативний настрій і 0 означає, що текст є нейтральним. Якість класифікації багато в чому залежить від якості словника.

Словники можуть бути створені з використанням різних методів [2]:

– Вручну побудовані словники [3] (простий, але не дуже швидкий метод). Наприклад General Inquire [4], який складається зі слів суспільствознавчих категорій для контент-аналізу. Ці категорії аналізу контенту намагаються охопити тон, ставлення, зовнішній вигляд.

– Словники з підготовлених даних бувають напівавтоматичними (наприклад, використовують такі ресурси, як WordNet [5] або UNL [6]), або автоматичними, коли словник може бути отриманий автоматично через асоціацію, де оцінка для кожного нового прикметника розраховується з використанням частоти близькості від прикметника до одного або більшої кількості затравочних слів.

Як зазначено в [9], *методи машинного навчання для аналізу текстів* – це сукупність методів, заснованих на алгоритмах штучного інтелекту, які використовують для навчання дані раніше помічені як позитивні, негативні або нейтральні.

У спрощеному вигляді, задача класифікації текстів може бути описана наступним чином – задано набір маркованих даних:

$$T_{data} = \{(t_i, L_i),...(T, n)\}, \quad (2)$$

де кожен текст належить до набору даних T і мітка L_i є попередньо встановленим класом всередині групи класів L, мета полягає в тому, щоб побудувати алгоритм навчання, який буде приймати в якості вхідних даних навчальний набір T_{data} і створити модель, яка буде точно класифікувати немарковані тексти t_i у кількості n.

Найпопулярніші алгоритми навчання для класифікації тексту – це метод опорних векторів [7], наївний класифікатор Баєса [8] , дерева прийняття рішень [9, 10], метод максимальної ентропії [11] та нейронні мережі [12].

Метод опорних векторів (SVM) [7] – це метод навчання з учителем, що використовується для бінарної класифікації. Даний алгоритм машинного навчання будує розділяючу поверхню у гіперпросторі з точок (об'єктів вибірки), що лежать між полярними підмножинами, тобто розмежовує класи. Точки побудованої поверхні називаються опорними векторами. Цей класифікатор може замінювати нейронні мережі, але має дуже повільний процес навчання.

Знаходження параметрів SVM відповідає опуклій оптимізації. Завдання класифікації, як правило, включає в себе поділ даних на навчальні та тестові набори. Кожен екземпляр в навчальному наборі містить одне «цільове значення» (тобто клас-мітку) і кілька «атрибутів» (функції спостереження за змінними). Метою SVM є вироблення моделі (на основі навчальних даних), яка визначає цільове значення тексту, та побудова оптимальної розмежувальної гіперплощини. SVM для класифікації використовується, щоб знайти лінійну модель такого вигляду:

$$y(x) = \omega^{\gamma} x + b, \quad (3)$$

де x вхідний вектор, ω і b є параметрами, які можуть бути скориговані для певної моделі, що оцінюється емпіричним шляхом, γ – вектор двоїстих змінних.

Для простої лінійної класифікації завдання полягає в тому, щоб звести до мінімуму функцію помилок, що визначається рівнянням:

$$C\sum_{n=1}^{N} \xi_n + \frac{1}{2}\|\omega\|^2 \to \min, \quad (4)$$

де C – обрана константа;

ω – вектор коефіцієнтів;

ξ – параметр для обробки неподільних даних (входів);

n – номер процедури навчання.

Дерева рішень [9, 10] можуть бути адаптовані до практично будь-якого типу даних, тому цей спосіб широко використовується в алгоритмах машинного навчання. При контрольованому машинному навчанні використовується алгоритм, який ділить підготовлені дані на більш дрібні частини, з метою визначення моделі, яка може бути використана для класифікації. Дані потім представляються у вигляді логічних структур, подібних до древовидної, які можуть бути легко зрозумілі без будь-яких статистичних знань. Алгоритм особливо добре підходить для випадків, коли може бути знайдено багато ієрархічних категоріальних відмінностей. Вони побудовані з використанням евристичних алгоритмів, які називають рекурсивним розбиттям. Це, як правило, відомо, як підхід «розділяй і володарюй», оскільки він використовує значення функцій для поділу даних на менші підмножини подібних класів. Структура дерева рішень

складається з кореневого вузла, який представляє собою весь набір даних, рішень вузлів, які виконують обчислення і листових вузлів, які здійснюють класифікацію.

Для того, щоб класифікувати невідомий екземпляр, дані передаються через дерево. На кожному вузлі рішення певної функції, отриманої з вхідних даних, порівнюються з константою, яку було визначено на етапі підготовки. Обчислення, яке відбувається в кожному вузлі рішення, зазвичай, порівнює обрану функцію з цією, заздалегідь заданою, константою, тоді рішення буде ґрунтуватися на функції, створюючи два способи поділу на дереві. Дані будуть, в кінцевому підсумку, проходити через ці вузли рішення до тих пір, доки не досягнуть листового вузла, який представляє собою визначений клас. Є багато різних реалізацій і варіантів алгоритму дерев рішень, такі як, Random Forest та метод J48, який є реалізацією Java алгоритму C4.5. Для виявлення спаму, пропаганди, інформаційно-психологічних впливів дерева прийняття рішень на практиці застосувати неможливо, оскільки вони не підтримують інкриментого навчання. Можна взяти великий об’єм даних та побудувати для нього дерево рішень, але врахувати нові повідомлення даний метод не зможе, оскільки для нових даних його доведеться кожен раз навчати заново.

В якості найпростішого методу для класифікації тональності тексту, а також фільтрації спаму, використовується *наївний класифікатор Баєса* [8]. У даному класифікаторі використовується теорема Баєса для визначення ймовірності приналежності елемента вибірки до одного з класів при припущенні незалежності ознак. Для підвищення якості класифікації застосовується метод *максимальної ентропії* [11]. Класифікатор максимальної ентропії є класифікатором ймовірності, який належить до класу експоненціальної моделі. На відміну від наївного класифікатора Баєса, він не припускає, що ознаки умовно незалежні одна від одної. Цей класифікатор засновано на принципі максимальної ентропії усіх моделей, які відповідають даними навчання з найбільш рівномірним розподілом. Класифікатор максимальної ентропії може бути використаний для вирішення великої кількості різноманітність завдань класифікації тексту, таких як виявлення спаму, сленгу, тематичної класифікації тощо.

Для автоматичного визначення емоційного забарвлення контенту соціальних мереж можна використовувати *нейронні мережі* [12]. За допомогою нейронних мереж можна проводити аналіз емоційного навантаження смайлів та картинок, що постяться у соціальних мережах, а також аналіз емоційного забарвлення текстових даних.

Перевагами застосування нейронних мереж є: можливість рішення задач при невідомих закономірностях, здатність до навчання, стійкість до шумів у вхідних даних.

Алгоритми навчання штучних нейронних мереж поділяються на алгоритми навчання з учителем та без учителя. Навчання нейронної мережі в першу чергу полягає в зміні вагових коефіцієнтів синаптичних зв'язків між нейронами.

Для аналізу текстових даних, доцільно застосовувати глибоке навчання рекурентних нейронних мереж, яке не викликає складнощів із перенавчанням, на відміну від згороточних та повнозв’язних нейромереж.

Переваги та недоліки досліджених методів автоматичного визначення тональності контенту у віртуальних соціальних мережах представлені у таблиці 1.

Таблиця 1 – Порівняльний аналіз методів автоматичного визначення тональності у текстах та мультимедійних даних

Назва методу	Необхідність застосування словників	Необхідність попередньої лінгвістичної обробки тексту	Можливість застосування до різних типів даних	Можливість застосування для виявлення інформаційно-психологічних впливів
Лексемний метод	+	+	Тільки до текстів	+
Метод опорних векторів	–	–	До різних типів даних	+
Дерева прийняття рішень	–	+	До різних типів даних	–
Наївний класифікатор Баєса	+	–	Тільки до текстів	+
Метод максимальної ентропії	+	–	Тільки до текстів	+
Нейронні мережі	–	–	До різних типів даних	+++

Висновки. Дослідження методів автоматичного аналізу настроїв в соціальних мережах показало, що найбільш придатними для виявлення у текстах та мультимедійних даних інформаційно-психологічних впливів, заснованих на емотивній комунікації, є нейронні мережі, оскільки вони не потребують складання словників, обов'язкової попередньої лінгвістичної обробки текстів, можуть застосовуватися до різних типів даних та здатні здійснювати класифікацію за декількома категоріями, що дозволить виявляти різні типи інформаційно-психологічного впливу. Також для поставленої задачі можна використовувати лексемний метод, метод опорних векторів, наївний класифікатор Баєса та метод максимальної ентропії. Недоліком методу опорних векторів є те, що він здійснює бінарну класифікацію, яка дозволить розділити дані тільки на дві категорії: дані без інформаційно-психологічних впливів та дані з інформаційно-психологічними впливами. Основний недолік наївного класифікатору Баєса – неможливість врахування залежності результату від комбінації ознак (слів). Спільним недоліком лексемного методу, наївного класифікатору Баєса та методу максимальної ентропії є необхідність складання словників, що вимагає тісної співпраці з лінгвістами.

Слід зазначити, що будь-який з методів автоматичної класифікації тексту не може дати беззаперечних результатів. Помилки даних методів пояснюються наступними проблемами: орфографічними помилками у тексті, відсутністю зв'язків у тексті. Також не завжди про ставлення автора можна сказати за наявністю або відсутністю позитивних, негативних або нейтральних відгуків.

Для поліпшення автоматичного визначення тональності тексту доцільно використовувати декілька систем класифікації, гібридні методи класифікації. Також, для підвищення якості роботи класифікаторів треба забезпечити автоматичне виправлення орфографічних помилок, вдосконалювати словники (для методів,

заснованих на словниках) і навчальні вибірки (для методів машинного навчання). Також можна підвищити точність роботи алгоритмів автоматичної класифікації тексту, застосовуючи існуючі розробки з автоматичного реферування, виявлення кореферентності тощо.

Список літератури

1. Stone, P. A computer approach to content analysis: Studies using the general inquirer system [Text] / P. Stone, E. Hunt // Spring Joint Computer Conference, AFIPS '63 (Spring) / P. Stone, E. Hunt. – New York: ACM, 1963. – С. 241–256.
2. Lexicon-based methods for sentiment analysis [Text] / [M. Taboada, J. Brooke, M. Tofiloski та ін.]. // Computational linguistics. – №37. – С. 267–307.
3. Wiebe, J. Learning subjective adjectives from corpora [Text] / J. Wiebe // In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence / J. Wiebe., 2000. – С. 735–740.
4. Guthrie, D. Unsupervised detection of anomalous text. [Text] : дис. докт. філос. наук / Guthrie D. – Sheffield, 2008. – 186 с.
5. Hayashi, Y. A Dictionary Model forUnifying Machine Readable Dictionaries and Computational Concept Lexicons [Text] / Y. Hayashi, T. Ishida // LREC / Y. Hayashi, T. Ishida., 2006. – С. 1–6.
6. Boudhh, S. Unification of universal word dictionaries using WordNet ontology and similarity measures [Text] / S. Boudhh, P. Bhattacharyya, // In Proceedings of the 5th International Conference on Global WordNet / S. Boudhh, P. Bhattacharyya,. – Mumbai, 2010. – С. 1–7.
7. Cortes, C. Support-vector networks. [Text] / C. Cortes, V. Vapnik // In Machine Learning / C. Cortes, V. Vapnik. – Hingham: Kluwer Academic Publishers, 1995. – (volume 20). – С. 273–297.
8. Narayanan, V. Fast and accurate sentiment classification using an enhanced naive bayes model. [Text] / V. Narayanan, I. Arora, A. Bhatia // Intelligent Data Engineering and Automated Learning IDEAL / V. Narayanan, I. Arora, A. Bhatia. – Berlin: Springer, 2013. – (volume 8206 of Lecture Notes in Computer Science). – С. 194–201.
9. Mitchell, T. Machine Learning [Text] / T. M. Mitchell. – New York: McGraw-Hill, 1997. – 414 с.
10. Schmid, H. Probabilistic part-of-speech tagging using decision trees [Text] / H. Schmid // In International Conference on New Methods in Language Processing / H. Schmid. – Manchester, 1994. – С. 44–49.
11. Patel, D. Sentiment Analysis using Maximum Entropy Algorithm in Big Data. [Text] / D. Patel, S. Saxena, T. Verma. // International Journal of Innovative Research in Science, Engineering and Technology. – 2016. – №5. – С. 8355–8361.
12. Vyrva, N. Sentiment Analysis in Social Media. Master's Thesis in Computer Science [Text] / N. Vyrva. – Halden: Ostfold University College, 2016. – 82 с.

Dmitry Shyngalov, post-graduate, Yelyzaveta Meleshko, Assoc. Prof., Phd tech. sci., Roman Mynaylenko, Assoc. Prof., Phd tech. sci., Vitaliy Reznichenko, Lect.
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine

Methods of automated sentiment analysis on social networks

The article analyzes existing methods of automatic analysis of sentiments in online virtual social networks, for to identify methods that can determine the information-psychological influences, based on emotive communication.

The analysis includes a review of lexical method and methods of machine learning. Specifically there are considered procedure of preprocessing text before the analysis and categories of dictionaries used for analysis. The algorithms for classification study the text are given: the method of support vectors, naive Bayesian classifier, decision tree method of maximum entropy, and neural networks.

Methods to improve the quality of results of automatic determination of the tone of the text are proposed.

social networks, sentiment analysis, lexical analysis, machine learning

Одержано 15.05.17