УДК 32.973.202:07.681

д.фіз.-мат.н., доц. **Марченко О.О.** (КНУ)
к.т.н. **Марченко-Бабіч О.М.** (ВІКНУ)

БАГАТОМІРНА МОДЕЛЬ ОЦІНЮВАННЯ СЕМАНТИЧНОЇ ЗАБАРВЛЕНОСТІ ПРИРОДНОМОВНИХ ТЕКСТІВ

У роботі описується розробка системи класифікації емоційного сприйняття природномовних текстів, яка включає напрацювання у галузі семантичного аналізу, лінгвістики і когнітивної психології. Пропонується багатомірна модель розмітки і оцінювання семантичної забарвленості текстів природною мовою, у якій враховується широкий спектр людських емоцій.

Наведено результати опрацювання розміченого експертами текстового корпусу, на якому в подальшому здійснювалось навчання системи, що була реалізована на основі моделей машинного навчання. Для побудови системи мультисентимент-класифікації текстів були використані Модель наївного Байєсівського класифікатора (BNM), Модель лінійної регресії (LRM) і Метод опорних векторів (SVM) у різних варіантах і конфігураціях.

Наводяться дані експериментів, проведених для перевірки ефективності і коректності моделі. Вони включають навчання та перевірку точності сентимент-класифікаторів, з використанням методу перехресного оцінювання (так званого методу крос-валідації – cross-validation method).

Запропонована модель відрізняється від стандартних підходів технології сентимент-аналізу, і дозволяє ідентифікувати широкий набір базових емоцій людини, що збільшує можливості визначення семантичної забарвленості тексту із подальшим використанням для вирішення завдань інформаційно-аналітичної діяльності в інтересах інформаційної безпеки. Результати проведених експериментів підтверджують ефективність і коректність розробленої моделі.

Ключові слова: система, сентимент-аналіз, семантична забарвленість, модель машинного навчання, класифікація.

Вступ. Сентимент-аналіз текстів є одним з найпопулярніших та потрібних класичних завдань обробки природної мови. Наукові дослідження, які проводились протягом десятиліть в цій галузі, сприяли створенню низки сучасних систем сентимент-аналізу з різними принципами роботи. Переважно вони базуються на представленні векторного простору на рівні речень. Так, у [1] є підхід до відслідкування розподілу емоцій у тексті без використання парсерів і ресурсів лексики на позначення почуттів. У моделі [2] вивчається композиційне векторне представлення для фраз та речень довільної довжини і досліджуються вектори слова і те, як вони впливають на сусідні слова. Поєднання нової моделі і даних (рекурсивних нейронних тензорних мереж і набору дерев) у [3] реалізовано у системі сентимент-аналізу по оцінюванню цілого речення. Використання системи сентимент-аналізу для мережі “Twitter” [4] в реальному часі із залученням розроблених експертами правил і ключових слів дозволяє отримати повну і точну картину політичної обстановки, наявної в конкретний час в режимі online. Варто вказати нові підходи до створення систем сентімент-аналізу, що включають розробку послідовності різноманітних ознак для повідомлень, які надсилаються у “Twitter”. Набір даних, отриманих з громадських висловлювань, синтаксичні особливості, кластери і вкладення служать для навчання класифікатору методу опорних векторів. При цьому використовуються методи навчання зі вчителем, без вчителя або частково зі вчителем [5]. Нещодавні наукові дослідження включають визначення рівня агресивності тексту шляхом аналізу позитивного і негативного впливів окремих слів на змістовне значення тексту в цілому та оцінювання його звуко-кольорових характеристик, а також шляхом оцінювання емоційного впливу фонетичної структури текстів на підсвідомість людини [6].

Усе це робить можливим застосовувати такі системи у різних сферах суспільного життя, у тому числі в галузі інформаційної безпеки. Вони розширюють можливості інформаційно-аналітичних структур щодо опрацювання важливих показників аналізу інформаційного простору. Рішення, запропоновані даними системами, дозволяють деталізувати загальний методичний апарат, який розглядається у вітчизняних наукових

розробках [7- 8], в якому процес аналізу важливих показників інформаційного простору сприяє використанню існуючих напрацювань в інтересах інформаційної безпеки.

У вищезгаданих джерелах у класичній постановці завдання йдеться про класифікацію фрагменту текста щодо його емоційної складової, тобто як до даного фрагменту ставиться автор текста або які емоції він викликає у читача. Зазвичай оцінювання фрагментів текста здійснюється по шкалі «негатив-нейтрал-позитив». Більш точне ідентифікування емоцій є більш важким завданням, що вимагає більш складної обробки текста, включно до реалізації елементів семантичного аналізу, а тому розробка систем мультисентимент-аналізу була скоріше метою і завданням наукових досліджень, ніж реальних науково-дослідних проектів для створення прикладних систем практичного застосування. Тому, **метою даної статті** є представлення основних положень розробки багатомірної моделі оцінювання семантичної забарвленості природномовних текстів.

У розробленій авторами моделі враховується широкий спектр людських емоцій, таких як радість, повага, емпатія, страх, ненависть тощо.

Згідно створеної моделі, експертами було розмічено корпус текстів, на якому в подальшому здійснювалось навчання системи, реалізованої на основі моделей машинного навчання. Для побудови системи мультисентимент-класифікації текстів були використані Модель наївного Байєсівського класифікатора (BNM), Модель лінійної регресії (LRM) і Метод опорних векторів (SVM) у різних варіантах і конфігураціях.

Для перевірки ефективності і коректності моделі було проведено ряд експериментів навчання та перевірки точності сентимент-класифікаторів, з використанням методу перехресного оцінювання (так званого методу крос-валідації – cross-validation method).

Розробка набору еталонних емоцій

При створенні текстового корпусу для моделі оцінювання враховано, що будь-який стан людини містить у собі певні емоції як невід'ємну складову [9]. Найпотужніші емоції виникають в початковий період формування стану як суб'єктивні реакції людини, яка виражає своє відношення до процесу реалізації актуальної потреби. Для розпізнавання визначено такі групи семантично-забарвлених лексичних одиниць:

- назви емоцій та почуттів;
- перелік психічних станів, які є супутніми у сприйнятті людини.

До лінгвістичного забезпечення блоку сентимент-аналізу включено напрацювання когнітивної психології, яка вивчає когнітивні процеси та особливості функціонування психіки при сприйнятті нової інформації [10].

Брались до уваги такі властивості психіки як динамічність, рухливість, схильність до вживання стереотипів та когнітивна, афективна і соціальна функції [9].

На підставі вивчених даних у якості базових емоцій для мультисентимент-класифікаторів були вибрані градації літерами латинського алфавіту від A (агресія, ненависть) до V (пригніченість).

Створення корпусу розмічених текстів

Для отримання навчального набору текстів для створення класифікаторів семантичної забарвленості було розмічено корпус англомовних статей політичної спрямованості. Він містить близько 1000 текстів, узятих з таких видань як The Guardian, The BBC, The telegraph, Euronews, France24, Deutschewelle, Tass, Russia today, Sputnik, Bloomberg, The CNN, The New York Times.

Для розмітки використовувався XML-подібний код. Кожен текст корпусу маркувався декількома експертами, відбирались ті фрагменти, що мали співпадіння по межах та по назві.

Для побудови класифікаторів семантичної забарвленості у текстах були використані такі моделі машинного навчання:

- модель наївного Байєсівського класифікатора (BNM);
- метод лінійної регресії (LRM);
- метод навчання зі вчителем (svm with linear trainer);
- метод навчання з радіальним базисним ядром (svm with trainer with radial basis kernel);

- метод навчання з ядром перетину гістограм (svm with trainer with histogram intersection kernel).

Для їх реалізації було використано бібліотеку sklearn.py для мови програмування python.

Набір ознак для машинного навчання класифікаторів

Для навчання класифікаторів були реалізовані наступні ідентифікатори особливостей:

1. Уніграми
2. Біграмми
3. Триграми
4. Уніграми з синонімів оригінальних уніграм
5. Біграми з синонімів оригінальних біграм
6. Триграми з синонімів оригінальних триграм
7. Уніграми з найближчих сусідів по синсетах WordNet
8. Біграми з найближчих сусідів по синсетах WordNet
9. Триграми з найближчих сусідів по синсетах WordNet

Міри семантичної близкості, розроблені в моделях машинного перекладу, використовуються також:

10. BLEU [11]

$$BLEU(r,c) = BP(r,c) \times exp\left[\sum_{n=1}^{N} \frac{1}{N} \times \log\left(p_n(r,c)\right)\right], \quad (1)$$

де N – *максимальний розмір* n-грами. Тоді точність p_n визначається так:

$$p_n(r,c) = \frac{\sum_{x \in NGrams_n(c)} count\left(x, NGrams_n(r) \cap NGrams_n(c)\right)}{\sum_{x \in NGrams_n(c)} count\left(x, NGrams_n(c)\right)}, \quad (2)$$

де значення (x, X) – значення елементу x в множині X.

11. BLEU, де послідовності значущих слів використовуються для N-грам, що відповідає:

$$IDF(x, docs) > L, \quad (3)$$

де IDF – абревіатура для зворотної частоти присутності документу в наборі серед інших документів.

$$IDF(x, docs) = log\left(\frac{|docs|}{|docs|_{x \in docs|}}\right); \quad (4)$$

docs – корпус документів; L – деякий поріг, який залежить від особливостей корпусу документів.

12. BLEU, де використовуються послідовності синтаксичних N-грам.
13. NIST [12]
14. METEOR [13]
15. BADGER [14].

Експерименти

Експерименти з перевірки ефективності і коректності побудованих класифікаторів були проведені за принципом крос-валідації, коли навчання проводиться на одній частині розміченого тексту, а перевірка точності знаходження фрагментів, що містять емоції/психічні стани, і визначення конкретного класу здійснюється на іншій частині корпусу, що не була задіяна при навчанні. Тоді отримані результати можна порівняти з

еталонними розмітками. Далі частини, що навчають та перевіряють, міняються місцями, і процес навчання і перевірки виконується знову. З отриманих оцінок точності і повноти знайдених рішень вибираються мінімальні значення як гарантовані. Отримані результати по набору емоцій та по реалізованих класифікаторах продемонстровано у табл. 1.

Таблиця 1

Результати експериментів

Літера	Кількість знаходжень у тексті 25492	Лінійний SVM	SVM з радіальним базисним ядром	SVM з ядром перетину гістограм	Лінійний SVM з компонентами WordNet
A	99	62,35%	61,12%	73,29%	
B	76	66,36%	66,79%		45,38%
C	319	0,00%	49,99%	62,24%	
E	487	0,05%	8,76%		1,32%
F	35	28,57%	62,32%		0,00%
G	72	12,50%	63,47%		14,27%
H	61	74,97%	57,21%	72,70%	
I	60	66,04%	57,53%	66,20%	
J	191	53,61%	62,24%		29,45%
K	38	69,35%		70,64%	19,76%
L	228	57,42%	55,30%	61,54%	
M	84	58,56%	64,12%	66,66%	
N	57	67,69%	56,43%	68,73%	
O	35	78,63%	67,66%	72,70%	
P	70	70,29%	63,21%	64,06%	
Q	43		65,15%	72,85%	29,58%
R	33	68,77%	64,37%	80,21%	
S	82	63,86%	62,56%	63,70%	
T	51	69,82%		71,66%	35,37%
U	157	60,09%	50,97%		28,60%
V	52	58,37%	55,33%		64,30%

Як видно з отриманих даних, різні класифікатори по-різному опрацьовують набори емоцій/психічних станів. В подальших дослідженнях, очевидно, доцільним є використання дворівневого ансамблю класифікаторів, коли до реалізованих класифікаторів першого рівня додається суперкласифікатор, який буде навченим на основі розв'язків класифікаторів першого рівня обчислювати кінцевий розв'язок.

Висновки. Таким чином, запропонована багатомірна модель оцінювання семантичної забарвленості речень тексту є основною відмінністю від стандартних підходів в сентимент-аналізі, і включає досягнення у сферах семантичного аналізу, лінгвістики і когнітивної психології. Результати проведених експериментів підтверджують ефективність і коректність розробленої моделі. Її використання дає можливість ідентифікувати широкий набір базових емоцій та психічних станів людини в системі класифікації емоційного сприйняття текстів природною мовою, більш точно визначити семантичну забарвленість тексту. Отримані результати доцільно використовувати в інформаційно-аналітичній діяльності для аналізу показників інформаційного простору. Саме врахування його важливих характеристик сприяє вирішенню завдань із забезпечення інформаційної безпеки.

ЛІТЕРАТУРА:

1. Socher R., Pennington J., H.E.H.N.A.Y., D., M.C.: Semi-supervised recursive autoencodersfor predicting sentiment distributions. In: Proceedings of EMNLP (2011).

2. Socher R., Huval B., M.C.D., Y., N.A.: Semantic compositionality through recursive matrixvector spaces (2012).

3. Socher R., Perelygin A.,W.J.Y.C.J.M.C.D.N.A.Y., P., P.C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of EMNLP (2013).

4. Wang H., Can D., K.A.B.F., S., N.: A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In: Proceedings of ACL (2012).

5. Jabreel M., Moreno A. Task 4: Sentiment Analysis in Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluations (2017).

6. Алімпієв А.М. Теоретичні основи створення технологій протидії прихованим інформаційним атакам в сучасній гібридній війні / А.М.Алімпієв, В.В. Бараннік, Т.В. Белікова, С.О. Сідченко // Режим доступу: www.hups.mil.gov.ua/periodic-app/article/17669/soi_2017_4_26.pdf

7. Косогов О.М. Методика визначення пріоритетів показників, що характеризують рівень загроз інформаційній безпеці держави. – Х.: Збірник наукових праць Харківського університету Повітряних Сил, № 2 (39), 2014. – С.163-166.

8. Левченко О.В. Методика виявлення заходів негативного інформаційного впливу на основі аналізу відкритих джерел / О.В.Левченко, О.М.Косогов // Системи обробки інформації. – 2016. – №1(138). – С.100-102.

9. И.Г.Малкина-Пых. Гендерная терапия. Справочник практического психолога, 2003. Режим доступу: http://medicinapediya.ru/gendernaya-psihologiya_789/gendernyie-stereotipyi-41837.html

10. Н.И. Козлов. Когнитивная психология. Режим доступу: http://www.psychologos.ru/articles/view/kognitivnaya-psihologiya

11. Kishore Papineni, Salim Roukos, T.W., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of ACL (2002)

12. Doddington, G.: Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In: Proceedings of HLT. pp. 138–145 (2002)

13. Denkowski, M., Lavie, A.: Extending the meteor machine translation metric to the phrase level. In: Proceedings of NAACL (2010)

14. Parker, S.: Badger: A new machine translation metric. In: Proceedings of the Workshop on Metrics for Machine Translation at AMTA (2008).

REFERENCES:

1. Socher R., Pennington J., H.E.H.N.A.Y., D., M.C.: Semi-supervised recursive autoencodersfor predicting sentiment distributions. In: Proceedings of EMNLP (2011).

2. Socher R., Huval B., M.C.D., Y., N.A.: Semantic compositionality through recursive matrixvector spaces (2012).

3. Socher R., Perelygin A.,W.J.Y.C.J.M.C.D.N.A.Y., P., P.C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of EMNLP (2013).

4. Wang H., Can D., K.A.B.F., S., N.: A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In: Proceedings of ACL (2012).

5. Jabreel M., Moreno A. Task 4: Sentiment Analysis in Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluations (2017).

6. Alimpiyev A.M. Teoretychni osnovy stvorennja tehnologij protydiї prihovanym informacijnym atakam v suchasnij gibrydnij vijni / A.M.Alimpiєv, V.V. Barannik, T.V. Belikova, S.O. Sidchenko// Rezhim dostupu: www.hups.mil.gov.ua/periodic-app/article/17669/soi_2017_4_26.pdf

7. Kosogov O.M. Metodika viznachennja prioritetiv pokaznikiv, scho harakterizujut' riven' zagroz informacijnij bezpeci derzhavy. – H.: Zbirnik naukovih prac' Harkivs'kogo universitetu Povitrjanih Sil, № 2 (39), 2014. – S.163-166.

8. Levchenko O.V. Metodika vijavlennja zahodiv negativnogo informacijnogo vplyvu na osnovi analizu vidkrytyh dzherel / O.V.Levchenko, O.M.Kosogov//Sistemy obrobki informacії. – 2016. – №1(138). – S.100-102.

9. I. G. Malkina-Pyh. Gendernaja terapija. Spravochnik prakticheskogo psihologa, 2003. Rezhim dostupu: http://medicinapediya.ru/gendernaya-psihologiya_789/gendernyie-stereotipyi-41837.html.

10. Kozlov N.I. Kognitivnaja psihologija. Rezhim dostupu:

http://www.psychologos.ru/articles/view/kognitivnaya-psihologiya
11. Kishore Papineni, Salim Roukos, T.W., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of ACL (2002)
12. Doddington, G.: Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In: Proceedings of HLT. pp. 138–145 (2002)
13. Denkowski, M., Lavie, A.: Extending the meteor machine translation metric to the phrase level. In: Proceedings of NAACL (2010)
14. Parker, S.: Badger: A new machine translation metric. In: Proceedings of the Workshop on Metrics for Machine Translation at AMTA (2008).

Рецензент: д.тех.н., проф. Замаруєва І.В., професор кафедри Державного університету телекомунікацій

д.ф.-мат.н., доц. Марченко А.А., к.т.н. Марченко-Бабич О.Н.

МНОГОМЕРНАЯ МОДЕЛЬ ОЦЕНИВАНИЯ СЕМАНТИЧЕСКОЙ ОКРАСКИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

В работе описывается разработка системы классификации эмоционального восприятия естественноязыковых текстов, в которую включены наработки в области семантического анализа, лингвистики и когнитивной психологии. Предлагается многомерная модель разметки и оценивания семантической окраски текстов на естественном языке, в которой включён широкий спектр человеческих эмоций.

Приводятся результаты обработки размеченного экспертами текстового корпуса, на котором в дальнейшем выполнялось обучение системы, реализованной на основе моделей машинного обучения. Для построения системы мультисентимент-классификации тестов были использованы Модель наивного Байесовского классификатора (BNM), Модель ленейной регресии (LRM) и Метод опорных векторов (SVM) в разных вариантах и конфигурациях.

Приводятся данные экспериментов, проведённых для проверки эффективности и корректности модели. В них включены обучение и проверка точности сентимент-классификаторов, и используется метод перекрёстного оценивания (так называемого метода кросс-валидации - cross-validation method).

Предложенная модель отличается от стандартных подходов технологии сентимент-анализа, и позволяет идентифицировать широкий набор базовых эмоций человека, что увеличивает возможности определения семантической окраски текста с дальнейшим использованием для решения задач информационно-аналитической деятельности в интересах информационной безопасности. Результаты проведённых экспериментов подтверждают эффективность и коректность разработанной модели.

Ключевые слова: система, сентимент-анализ, семантическая окраска, модель машинного обучения, класификация.

Prof. Marchenko O.O., Ph.D. Marchenko-Babich O.M.

MULTIDIMENSIONAL MODEL OF NATURAL LANGUAGE TEXTS SEMANTIC COLOURING ASSESSEMENT

The article describes the development of classification system of natural language texts emotional perception that encompasses semantic analysis, linguistics and cognitive psychology groundwork. Multidimensional emotional model for the text sentences semantic colouring estimate is exposed. It makes possible to identify a wide spectrum of basic human emotions that is the principle difference from standard approaches in the sentiment-analysis domain.

Text training set which consists of texts marked by experts is exposed. It makes the basis for the system training which was realised with machine learning models. Such techniques as Bayes Naïve model, Linear regression model, Support vectors machine were used for creation of multisentiment classification system.

Results of experiments for assessments of precision and recall for found solutions are presented. They include learning and check of precision for sentiment-classifiers, with employment of cross-validation method.

The proposed multidimensional emotional model for text sentences evaluation is the principal difference from standard approaches in the sentiment-analysis. Its exploitation makes possible to identify a vast scope of basic human emotions in the system for classification of natural language texts emotional perception. Further employment of the system enables information and analytical activity for the purpose of information security. The experimental results cited in this paper confirm the efficiency and correctness of the developed model.

Keywords: system, sentiment-analysis, semantic coloring, machine learning model, classification.