DOI: https://doi.org/10.32839/2304-5809/2020-11-87-98
УДК 81.33

Левченко О.П., Поворознік Н.М.
Національний університет «Львівська політехніка»

ОСОБЛИВОСТІ ЗАСТОСУВАННЯ СЕНТИМЕНТ-АНАЛІЗУ (ІНТЕРПРЕТАЦІЯ САРКАЗМУ, БАГАТОЗНАЧНОСТІ, ЗАПЕРЕЧЕННЯ ТА МУЛЬТИПОЛЯРНОСТІ)

Анотація. Дані із соціальних мереж привертають увагу дослідників завдяки широкому використанню їх у повсякденному житті та чиннику новизни у поєднанні з доступністю, що слугує сильною мотивацією для досліджень аналізу тональності текстів. З огляду на це, постала низка технічних проблем, які ще не вирішили ні лінгвісти, ні NLP-спільнота. Часто у своїх публікаціях користувачі вживають саркастичні висловлювання, багатозначні слова, а в межах одного судження можуть бути наявні як позитивні, так і негативні настрої. Також це стосується і заперечних часток, які не завжди вказують на негативну тональність. У цій статті розглянуто чотири виклики, з якими стикаються дослідники під час проведення сентимент-аналізу, а саме: сарказм, заперечення, багатозначність та мультиполярність. Ці аспекти значно впливають на точність результатів під час визначення тональності. Також висвітлено сучасні підходи до розв'язання питання.

Ключові слова: сентимент-аналіз, сарказм, заперечення, багатозначність, му-льтиполярність.

Levchenko Olena, Povoroznik Nataliia
Lviv Polytechnic National University

FEATURES OF SENTIMENT ANALYSIS IMPLEMENTATION (INTERPRETATION OF SARCASM, WORD AMBIGUITY, NEGATION, AND MULTIPOLARITY)

Summary. In the past decades, sentiment analysis has become one of the most active research areas in natural language processing, data mining, web mining, and information retrieval. The great demand in everyday life and the factor of novelty coupled with the availability of data from social networks have served as strong motivation for research on sentiment-analysis. A number of technical problems, most of which had not been attempted before, either in the NLP or linguistics communities have also generated strong research interests in academia. Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text. The entities can be products, services, organizations, individuals, events, issues, or topics. The field represents a large problem space. It improves not only the field of natural language processing but also management, political science, economics, and sociology because all these areas are related to the thoughts of consumers and public. User-generated content is full of opinions, because the main reason why people post messages on social media platforms is to express their views and opinions, and therefore sentiment analysis is at the centre of social media analysis. It turned out that user messages often contain plenty of sarcastic expressions and ambiguous words. Within one opinion both positive and negative sentiments can be present. This also applies to negative particles, which do not always indicate a negative tone. This article investigates four challenges faced by researchers while conducting sentiment analysis, namely: sarcasm, negation, word ambiguity, and multipolarity. These aspects significantly affect the accuracy of the results when we determine a sentiment. Modern ap-proaches to solving the problem are also covered. These are mainly machine learning methods, such as convolutional neural networks (CNN), deep neural networks (DNN), long short-term memory (LTSM), recurrent neural network (RNN), support vector machines (SVM), etc.

Keywords: sentiment analysis, sarcasm, negation, word ambiguity, multipolarity.

Постановка проблеми. Інтерес до чужої думки, мабуть, такий же давній, як і сама словесна комунікація, адже визначення того, що думають інші, завжди було важливою частиною збирання інформації та процесу ухвалення рішень [1]. У наш час активного розвитку інформаційних технологій люди використовують форуми, соціальні мережі, блоги та інші платформи, щоб ділитися своїми думками, тим самим генеруючи величезну кількість даних. Збирання інформації про згенерований користувачами контент вручну займає дуже багато часу. Саме тому з 2000-го року стрімко почав розвиватися й сентимент-аналіз. Він став однією з найбільш досліджуваних сфер опрацювання природної мови, глибинного аналізу даних та веб-майнінгу. Аналіз тональності текстів має широкий спектр застосування. Він вдосконалює не лише сферу опрацювання природної мови, а також постачає цінні дані для менеджменту, політології, економіки та соціології, адже всі ці сфери пов'язані з думками споживачів та громадськості [2]. Як і будь-який вид аналізу природної мови сентимент-аналіз має певні особливості. На перший погляд, проблема полягає лише у класифікації тексту, але якщо заглибитися в це питання, стає зрозуміло, що існує низка складних проблем, які впливають на точність сентимент-аналізу, як-от: автоматична ін-терпретація сарказму, різних типів заперечень, багатозначних слів та мультиполярності.

Аналіз останніх досліджень та публікацій. Дослідженням сентимент-аналізу, зокрема й особливостями його застосування займається багато закордонних науковців. У цьому дослідженні увагу зосереджено на працях: B. Pang, L. Lee, B. Liu, E. Camp, L. Kumar, A. Somani, P. Bhattacharyya, A. Ghosh, T. Veale, M. Dadvar, C. Hauff, F. de Jong, S. Pal, S. Ghosh, A. Nag, B. Agarwal, N. Mittal, P. Bansal, S. Garg, B. Wang, M. Liu.

Виділення не вирішених раніше частин загальної проблеми. На відміну від великого інтересу закордонних дослідників до проблем, що виникають під час застосування сентимент-аналізу, у вітчизняних працях їх розглядають та досліджують досить рідко. Оскільки ці чинники мають великий вплив на кінцевий результат аналізу тональності, в авторів виникла ідея виокремити та розглянути основні з них.

Мета статті. Мета цієї праці – розглянути кожну із проблем та зрозуміти, як вони впливають на якість класифікаторів настрою, а також з'ясувати те, які технології можна використати для вирішення тієї чи іншої задачі.

Виклад основного матеріалу. Безумовно, виникають труднощі з оцінюванням текстів, які містять слова в переносному значенні, адже його неможливо визначити за окремими компонентами вислову. Зростає інтерес, відповідно й кількість досліджень, присвячених розпізнаванню різноманітних тропів, особливо сарказму. У «саркастичному» тексті люди виражають негатив, використовуючи позитивно забарвленні слова. Це значно ускладнює процес розпізнавання сарказму сентимент-класифікаторами, якщо вони не спеціально розроблені враховувати цей аспект.

Найчастіше сарказм трапляється в контенті, створеному користувачами соціальних мереж, як-от: коментарі в Facebook, Twitter тощо. Дуже складно розпізнати сарказм без хорошого розуміння ситуації, конкретної теми та оточення. Це стосується не лише машин, а й людей. Різноманіття тем, інтересів, культур, знань про світ, які автори вживають у своїх коментарях, також ускладнюють успішне тренування моделей сентимент-аналізу.

Для початку, розгляньмо сарказм із погляду лінгвістики, у якій він широко досліджений. Авторка однієї з найбільш цитованих праць у цій галузі E. Camp пропонує виокремлювати такі типи сарказму:

1. Propositional: На перший погляд здається, що текст не саркастичний, проте він вміщує в собі неявний сарказм. Наприклад: «Схоже на чудовий план!».

2. Embedded: У тексті наявна невідповідність настроїв у формі самих слів та фраз. Наприклад: «Люблю, коли мене перебивають».

3. Like-prefixed: Текст передбачає заперечення зробленого аргументу. Наприклад: «Так, наче ті хлопці вірять у те, про що вони говорять».

4. Illocutionary: Невербальне мовлення (мова тіла, жести), що вказує на саркастичність висловлювання [3].

Дослідження Elisabeth Camp опубліковано у 2012 році. У 2017 році дослідники Стенфордського університету провели власне досить цікаве дослідження під назвою ««Having 2 hours to write a paper is fun!»: Detecting Sarcasm in Numerical Portions of Text», у якому йдеться про числовий сарказм. Цей тип сарказму дуже часто трапляється в соціальних мережах. Ідея, що стоїть за цим, пов'язана зі змінами числових значень, які потім впливають на полярність тексту. Наприклад:

1. «У цього телефона чудова батарея, заряд зберігається аж до 36 годин» (наявний сарказм).

2. «У цього телефона чудова батарея, заряд зберігається аж до 2 годин» (наявний сарказм).

3. «Надворі +30, мені так спекотно» (немає сарказму).

4. «Надворі –30, мені так спекотно» (наявний сарказм).

5. «Ми їдемо дуже повільно – лише 30 км/год» (немає сарказму).

6. «Ми їдемо дуже повільно – лише 165 км/год» (наявний сарказм).

Отже, ці речення відрізняються лише за використаним числом, це і є числовим сарказмом.

Існують різні підходи до автоматичного визначення сарказму, а саме:

– на основі правил (rule-based);

– статистичний (statistical);

– алгоритми машинного навчання (machine learning algorithms);

– глибоке навчання (deep learning) [6].

Підходи, засновані на глибинному навчанні, набувають все більшої популярності. L. Kumar, A. Somani, та P. Bhattacharyya у 2017 році дійшли висновку, що конкретна модель глибинного навчання (архітектура CNN-LSTM-FF) перевершує попередні підходи, досягаючи найвищого рівня точності виявлення числового сарказму. Однак глибинні нейронні мережі (deep neural networks) показали найкращі результати не лише під час визначення числового сарказму, вони також перевершили інші підходи до визначення сарказму загалом. A. Ghosh та T. Veale використовують комбінацію згорткової нейронної мережі (CNN), довгої короткочасної пам'яті (LSTM) та глибинних нейронних мереж (DNN). Свій підхід вони порівнюють із методом рекурсивних опорних векторів (SVM) та доходять висновку, що їхня архітектура глибинного навчання є більш досконалою [5; 6].

Другий аспект, який ускладнює сентимент-аналіз, це – заперечення. У лінгвістиці заперечення – це спосіб змінити полярність слів, фраз і навіть речень. Дослідники використовують різні лінгвістичні правила, щоб визначити наявність заперечення. Також важливо визначити діапазон слів, на які впливають заперечні слова. Немає усталеної кількості слів, які потрапляють під дію заперечення. Наприклад, у реченні *«Шоу було не цікавим»*, під дію заперечення потрапляє лише одне слово. Однак у реченнях на кшталт *«Я не вважаю цей фільм комедійним»*, заперечна частка *не* стосується всіх ужитих після неї слів.

Найпростіший підхід до роботи із запереченням у реченні, який використано в більшості сучасних методів сентимент-аналізу, це – маркування всіх слів як заперечних, починаючи від заперечного слова до наступного знаку пунктуації. Ефективність моделі може варіюватися через специфіку мовних конструкцій у різних контекстах.

Існує кілька форм висловлення негативної думки в реченнях:

1. Заперечення може бути морфологічним, коли воно утворюється за допомогою заперечних словотвірних афіксів, як-от: *не-, без-, обез-, зне-, ні, ані-, а-, анти-, ім, ін-, ір-, дис-, поза-, понад-, над-, недо-, небез-*.

2. Заперечення може бути імпліцитним, як у реченні *«З такою грою це буде його перша й остання вистава»* – воно передає негативні настрої, однак негативні слова не використовуються.

3. Заперечення може бути експліцитним, наприклад, *«Це не смачно»* [7].

Наявність вибірки різних типів описаних заперечень підвищить якість набору даних для навчання та тестування моделей сентимент-аналізу в межах заперечення. Згідно з останніми дослідженнями рекурентних нейронних мереж (RNN), різноманітні архітектури довгої короткочасної пам'яті (LSTM) перевершують усі інші підходи для виявлення типів заперечень у реченнях [8].

У статті «Effect of Negation in Sentiment Analysis» модель сентимент-аналізу оцінила 500 відгуків, які зібрали на сайтах Amazon та Trustedreviews.com. Автори порівнюють моделі з наявністю визначення заперечень та без неї. Їхній аналіз демонструє наскільки врахування чинника заперечення може підвищити точність моделі [7].

Ще однією проблемою, яка постає перед дослідниками сентимент-аналізу, є багатозначність слів. Вона полягає в неможливості заздалегідь визначити полярність певних слів, адже вона залежить від контексту. Серед наявних методів популярними є підходи сентимент-аналізу, засновані на словниках тональності. Словник вміщує слова з вказанням їхньої полярності. Деякі словники тональності доступні в інтернеті: SentiWordNet, General Inquirer, SenticNet та інші. Позаяк полярність слів відрізняється залежно від різних сфер, неможливо розробити універсальний лексикон, який подаватиме полярність кожного слова. Наприклад:

1. «Сюжет непередбачуваний».
2. «Це призвело до непередбачуваних витрат» [9].

Ці два приклади демонструють те, як контекст впливає на тональність судження. У першому прикладі, полярність слова *непередбачуваний* – позитивна. У другому прикладі те саме слово має негативну полярність.

Ще однією проблемою є мультиполярність. Інколи текст, який ми хотіли б проаналізувати, демонструє мультиполярність. У таких випадках наявність лише загального результату аналізу може ввести в оману. Уявімо, статтю чи рецензію, у якій мовиться про різних людей, продукти або компанії (чи їхні аспекти). Дуже часто в межах одного такого тексту висвітлюють і позитив, і негатив. У такому разі сумарній тональності бракує ключової інформації. Тому треба виділити в реченні всі сутності чи аспекти з мітками тональності й лише тоді підрахувати загальну полярність, якщо це необхідно. Розгляньмо приклад, який складається з декількох полярностей: «Їжа була смачною, а от обслуговування мені зовсім не до вподоби». Деякі моделі сентимент-аналізу визначать полярність, у цьому реченні як негативну або нейтральну. Для вирішення таких ситуацій, модель має визначати полярність кожного аспекту речення; тут *їжа* – це аспект позитивної полярності, а *обслуговування* – негативної [10]. Наприклад, у статті «Deep learning for aspect-based sentiment analysis», B. Wang та M. Liu зі Стенфордського університету пропонують підхід, що поєднує настрої з відповідними аспектами за допомогою синтаксичних дерев складових. Цей підхід продемонстрував конкурентоспроможність, а подекуди і кращу ефективність порівняно з найкращими результатами міжнародного семінару семантичної оцінки SemEval-2015 [10].

Висновки. Отже, з аналізу випливає, що дослідження сентимент-аналізу привертають увагу великої кількості науковців. Існує низка аспектів, які ускла-днюють його застосування. Найбільш поширені з них: сарказм, заперечення, багатозначність та мультиполярність. У дослідженнях сарказму найкращих результатів досягли за допомогою глибинних нейронних мереж, а також комбінації згорткових нейронних мереж (CNN) і довгої короткочасної пам'яті (LSTM). Популярним методом визначення заперечення є маркування всіх слів як заперечних, починаючи від заперечного слова до наступного знаку пунктуації. Однак на ефективність моделі впливають види конструкцій заперечення в різних контекстах, а щоб визначити тип заперечення найкраще використовувати архітектури довгої короткочасної пам'яті (LSTM). Щодо багатозначності, то поширеним є використання словників тональності, наприклад SentiWordNet, General Inquirer та інших. Належить зауважити, що такі словники не налаштовані на аналіз україномовного контенту. Під час аналізу тексту, де наявна мультиполярність для отримання більш точних результатів важливо звертати увагу на кожен аспект у реченні. Хороші показники можна отримати за допомогою сентимент-аналізу на основі аспектів. Урахування вищезгаданих особливостей значно підвищує точність моделей класифікації сентимент-аналізу.

Список літератури:

1. Pang B., Lee L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. Vol. 2. No. 1–2. 2008. Pp. 1–135.
2. Liu B. Sentiment analysis: Mining opinions, sentiments, and emotions, 2015.
3. Camp E. Sarcasm, pretense, and the semantics/pragmatics distinction. Noûs. 46. 2011.
4. Kumar L., Somani A., Bhattacharyya P. Approaches for computational sarcasm detection: A survey. *ACM CSUR*, 2017.
5. Ghosh A., Veale T. Fracking sarcasm using neural network. Proceedings of *NAACL-HLT 2016*. Association for Computational Linguistics, 2016.
6. Kumar L., Somani A., Bhattacharyya P. "Having 2 hours to write a paper is fun!": Detecting sarcasm in numerical portions of text, 2017.
7. Dadvar M., Hauff C., de Jong F. Scope of negation detection in sentiment analysis. *Dutch-Belgian Information Retrieval Workshop*, 2011.
8. Pal S., Ghosh S., Nag A. Sentiment analysis in the light of LSTM recurrent neural networks. *International Journal of Synthetic Emotions*. Pp. 33–39. 2018.
9. Agarwal B., Mittal N., Bansal P., Garg S. Sentiment analysis using common-sense and context information. *Computational intelligence and neuroscience*, 2015.
10. Wang B., Liu M. Deep Learning for aspect-based sentiment analysis. Stanford University report, 2015.

References:

1. Pang B., Lee L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135.
2. Liu B. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions.
3. Camp E. (2011). Sarcasm, pretense, and the semantics/pragmatics distinction. Noûs, 46.
4. Kumar L., Somani A., Bhattacharyya, P. (2017). Approaches for computational sarcasm detection: A survey. *ACM CSUR*.
5. Ghosh A., Veale T. (2016). Fracking sarcasm using neural network. Proceedings of *NAACL-HLT 2016*. Association for Computational Linguistics.
6. Kumar L., Somani A., Bhattacharyya P. (2017). "Having 2 hours to write a paper is fun!": Detecting sarcasm in numerical portions of text.
7. Dadvar M., Hauff, C., de Jong F. (2011). Scope of negation detection in sentiment analysis. *Dutch-Belgian Information Retrieval Workshop*.
8. Pal S., Ghosh S., Nag A. (2018). Sentiment analysis in the light of LSTM recurrent neural networks. *International Journal of Synthetic* Emotions, pp. 33–39.
9. Agarwal B., Mittal N., Bansal P., Garg S. (2015). Sentiment analysis using common-sense and context information. *Computational intelligence and neuroscience*.
10. Wang B., Liu M. (2015). Deep learning for aspect-based sentiment analysis. Stanford University report.