Лінгвістичні засади автоматичного сентимент-аналізу українськомовного тексту

Н. Дарчук

Інститут філології Київського національного університету імені Тараса Шевченка, Київ, Україна
Corresponding author. E-mail: nataliadarchuk@gmail.com.

Paper received 18.01.19; Accepted for publication 24.01.19.

https://doi.org/10.31174/SEND-Ph2019-189VII55-02

Анотація. У статті розглянуто лінгвістичні засади створення системи автоматичного аналізу тональності (sentiment analysis) публіцистичних та інформаційних повідомлень українськомовних джерел. Тональність – це забарвлення повідомлення лексикою, яка може впливати на свідомість об'єкта спілкування або спеціально нав'язуватися адресату. При цьому аналіз тональності покликаний відслідкувати не що говорять про той чи інший факт, подію, а наскільки емоційно це робиться. Визначено потенціал засобів реалізації сентименту в українській мові.

Ключові слова: *сентимент-аналіз., тональність, суб'єкт тональності, об'єкт тональності, аспект тональності, індекс тональності речення, індекс тональності тексту.*

Українська публіцистика, інформаційні повідомлення перебувають у постійному пошуку різних засобів увиразнення мовлення, щоб не тільки привернути увагу читача, глядача до якихось матеріалів, зацікавити його, а й вплинути на нього, приховуючи і факт впливу, і власні наміри аж до маніпуляції свідомістю людини.

Ці матеріали несуть не тільки інформаційне навантаження, а й сприяють формуванню суспільної думки своєю тональністю, яка, зокрема засобами емоційно-оцінної лексики, здатна навіювати такий психологічний стан у глядача поза його волею, який, як показали дослідження, потім важко піддається корекції [5].

Останнім часом інтерес дослідників до визначення тональності тексту, а саме визначення позитивного, негативного або нейтрального відношення автора до об'єктів, явищ, персон, згаданих у тексті, зріс у зв'язку із такими чинниками: ЗМІ та їхнє бажання будь-як привернути увагу як читачів, так і глядачів; Інтернет-спілкування, розвиток соціальних мереж – відгуки користувачів про сервіси, роботу компаній тощо [7]; особистісна характеристика автора за його текстами тощо. А оскільки стає можливим долучити до розв'язання, хоча б часткового, автоматичний аналіз тексту, ця проблема є активно обговорюваною у середовищі комп'ютерної лінгвістики, тому що зростає можливість автоматично вимірювати напругу у суспільстві – у соціологічних та політичних колах передбачати можливі результати виборів, переваги політичних партій тощо. Особливо актуальним є виявлення тональності тих текстів, які містять злобу, розповсюджують ненависть [5].

Мета даного проекту – створення системи автоматичного аналізу тональності (sentiment analysis) публіцистичних та інформаційних повідомлень українськомовних джерел. Під тональністю ми розуміємо забарвлення повідомлення лексикою, яка може впливати на свідомість об'єкта спілкування або спеціально нав'язуватися адресату. При цьому аналіз тональності покликаний відслідкувати не що говорять про той чи інший факт, подію, а наскільки емоційно це робиться. Тому аналіз тональності іноді називають сентимент-аналізом, аналізом емоційної складової повідомлень.

Спеціалісти в галузі соціальної психології стверджують, що свідомість людини схильна до спрощення і не схильна до протиріч, тому активно піддається впливу через штучну оману, часто перебільшує роль людських намірів і недооцінює об'єктивну дійсність. Інструментом мовленнєвого впливу, за А. Барановим, є використання мовних засобів, вплив яких іноді на перший погляд непомітний, але дуже значний та ігнорується адресатом повідомлення в угоду мовцеві [1, с. 213].

При аналізі тональності слід звернути увагу на чотири моменти:

- суб'єкт тональності (автор повідомлення);
- об'єкт тональності (про що йде мова);
- аспект тональності (характеристика об'єкта);
- оцінка тональности – сентимент [4].

Суб'єкт тональності може надавати характеристику об'єктам висловлювання цілісно в той чи інший спосіб і зводиться до бінарного (позитивного /негативного) представлення тональності, а дослідникові треба лише детальніше її покласифікувати, наприклад, за допомогою шкали оцінки. Об'єкт тональності – це зміст написаного, який може визначатися, наприклад, за ключовими словами. Аспект тональності – це характеристика, яка надається об'єкту тональності. А сентимент – це власне тональна оцінка – ставлення автора до описуваного предмета, аспекту, властивостей.

Велике значення для характеристики тональності має жанр текстів. Наприклад, в інформаційних (новинних) текстах можуть бути множинними як суб'єкти сентименту (автори та їхні оцінки), так і об'єкти, а також ситуації (аспекти), що значно ускладнює процес автоматичного аналізу тональності, як і багатозначність оцінної лексики, яка значною мірою залежить від контексту (*сталевий кулак* про фізичну силу боксера у спортивних новинах (позитивна оцінка) і *сталева конструкція* – нейтральна). Але тональність як явище міститься не в усіх типах текстів. Зупинимося на останньому компоненті тональності – оцінному.

Оскільки основною сферою, яка завжди привертала підвищену суспільну увагу, є політична комунікація, матеріалом нашого дослідження є як новинні стрічки, що транслюються телебаченням, так і статті у газетах різних регіонів України та інтернет-виданнях. Отже, це стандартизовані джерела ЗМІ, а також соціальні

медіа – записи у блогах, соціальних мережах – загалом корпус складає близько 10 млн слововживань. Зрозуміло, що вручну проаналізувати такий великий масив не можливо, тому було розроблено спеціальне програмне забезпечення, до якого входить база даних тональної лексики як лінгвістичного засобу аналізу українськомовного тексту.

Перед нами стояло завдання визначити потенціал засобів реалізації сентименту в українській мові..

Є два способи автоматичного оцінювання тональності: перший заснований на правилах, другий – на застосуванні методів машинного навчання, яке передбачає наявність великої колекції текстів із уже розміченими оцінками, на яких «тренуються» моделі машинного навчання [3]. Тобто оцінка значення тональності встановлюється дослідним шляхом, за допомогою розмічування операторами, яка потім використовується як «золотий стандарт» при сентимент-аналізі [6]. Оскільки для українськомовного тексту така система створюється вперше, використаємо перший спосіб – за правилами. Правилами на кшталт імплікативних (якщо… то) передбачена перевірка лексем або ланцюжків словоформ за спеціально укладеними словниками: якщо лексема належить до списку із позитивним сентиментом, то їй вона присвоюється. Серед правил, наприклад, перевірка на наявність заперечної частки *не*, яка алгоритмічно змінює оцінку на протилежну. З накопиченням ваг оцінок формується і сума ваг, і комбінація правил, що, на нашу думку, дає гнучку систему оцінювання.

Лінгвістичними компонентами системи є словники іменників, прикметників і прислівників, створені на матеріалі лексикону українськомовної публіцистичної лексики обсягом у 40 тис. лексем, відібрані за бінарною шкалою (позитив / негатив). Виділяють оцінний компонент як в денотативній, так і в конотативній частині тлумачення. Денотативна пов'язується з понятєвим ядром семантики слова, а емоційна – зі ставленням суб'єкта до об'єкта і міститься в конотативній частині. Оцінки встановлювалися поза контекстом трьома способами:

1) якщо у прямому значенні лексеми присутні семи з оцінним значенням (***вартість*** – позитивна якість, цінність; ***братання*** – вияв дружніх почуттів; vs: ***афера*** – ризикована справа, здійснювана з метою наживи, шахрайство; ***байдужість*** – відсутність почуттів);

2) якщо у переносному значенні присутні семи з оцінним значенням (***король*** – про що-небудь видатне в якому-небудь аспекті, що посідає помітне місце серед подібних; ***золото*** – ласкаве звертання до когось; vs: ***барліг*** – безладдя; ***болото*** – погане середовище, що справляє поганий вплив на кого-небудь);

3) якщо оцінне значення передається словотворчими засобами (*бараболь**ка, барвіноч**ок***; vs: *дів**аха, звір**юга***) (загальні дані див. у табл.1)

Табл.1. Співвідношення лексем з позитивною / негативною оцінкою

Частина мови	К-сть ЛСВ з позит. оцінкою	К-сть ЛСВ з негат. оцінкою	сума
Іменники (непредметні імена)	144	830	**974**
Іменники (предметні імена)	189	800	**989**
Прикметники	2260	3298	**5558**
прислівники	977	823	**1800**
Сума	**3570**	**5751**	**9321**

Як видно з таблиці, словник оцінної лексики складає 9321 ЛСВ. Найбільше в ньому іменників, негативна оцінка превалює над позитивною. Ці словники розмічені автоматизовано, не є вичерпними для української мови, оскільки охоплюють тільки лексику публіцистичного стилю. Бажано їх поповнювати лексикою інших стилів, жанрів та підмов.

Розмічування здійснювалося з урахуванням значення кожного ЛСВ. Такий словник спрямований на те, щоб автоматично здобувати з текстів оціночні слова й вирази і дослідити, які мовні засоби характерні для тієї чи іншої предметної галузі. Тестування показало, що одні оціночні вирази вживаються тільки в конкретних предметних галузях, інші є оціночними в одній і не є оціночними в іншій.

З логіко-філософської точки зору, оцінка є результатом пізнавальної діяльності людини в оточуючому середовищі, характер оцінки залежить від потреб, прагнень, цілей людини і базується на опозиції *хороший/поганий: хороший* – має відповідати ідеалізованій моделі світу, а *поганий* – не відповідає цій моделі. Основою оцінки є порівняння результатів людської діяльності з якісною сферою пізнавальної діяльності людини і як результат формування судження про ціннісні якості об'єкта судження. Виявляється, що є дві корелятивні сутності суб'єктно-об'єктних відношень: цінність характеризує об'єкт у його стосунках до суб'єкта, а оцінка – ставлення суб'єкта до об'єкта [4, с. 5]. Ми не будемо занурюватися в особливості логіко-семантичного, лінгвогносеологічного, функціонального трактування оцінки, нас цікавитиме мовний аспект, а саме: сукупність мовних одиниць, які несуть оцінне значення, виражають позитивне або негативне ставлення мовця до змісту мовлення.

Крім оцінки, яка закладена у денонативному значенні лексем (ЛСВ), до мовних засобів відносяться емоційно навантажена та експресивна лексика. З'ясуємо зміст понять «емоційність», «експресивність», «оцінка». Ми не розрізняємо «емоційну лексику», яка виділяється лише на основі функції вираження емоційного ставлення мовця до оточуючого, і назви емоції (*гнів, радість* тощо). Всі ці лексеми наділені особливою функцією вираження емоцій і протиставляються нейтральній лексиці. Крім того, до емотивних лексем включено широкий клас пестливих і лайливих слів, а також слова зі спеціальними афіксами, що мають емотивне значення, та емоційно-оцінні слова (*неук, брехун*). Щодо метафоричних найменувань, емотивність яких створюється за рахунок денотативного значення (про людину *слон, нарцис*), то вони не включені до пошукової бази через вторинність значення.

Під емоційністю в мові розуміють таку якість мови, що виражає почуття, переживання, настрої, суб'єктивне ставлення мовця до висловлюваного.

Експресивність (франц. expressif, від лат. expressus – виразний) – властивість мовної одиниці підсилювати логічний та емоційний зміст висловленого, виступати засобом суб'єктивного увиразнення мови. Через експресивність виражальних засобів мовець передає своє ставлення і до повідомлення, і до адресата [6, с. 156].

Оцінка – результат процесу оцінювання, взаємодії дійсності та людини у найрізноманітніших аспектах. Об'єктивний світ розчленовується людиною з точки зору його оцінного характеру – добра і зла, користі й шкідливості тощо, і це розчленування є соціально обумовленим і дуже складним чином відображено у мовних структурах [4].

Можна запропонувати такі типи оцінювальної лексики у сполученні з емоційною лексикою:

1) Слова з неоцінним денотативним компонентом – не оцінне, не емоційне (*стіл* – вид меблів);

2) Слова з оцінним денотативним компонентом: оцінне, не емоційне (*злодій* – той, хто займається злодійством; *артизм* – висока майстерність; *благозвучність* – приємне звучання);

3) Слова з неоцінним денотативним компонентом – неоцінне, емоційне (*благовірний* – чоловік, *бабій* – той, хто упадає коло жінок; залицяльник, зальотник)

4) Слова з оцінним денотативним компонентом – оцінне, емоційне (*віртуоз* – людина, що досягла найвищого ступеня майстерності в якійсь справі; *бидло* – груба, підла людина).

Ці типи становлять основу тонального шкалування: лексеми першого типу одержують оцінку «0», другого – «плюс vs мінус 1», третього – «плюс vs мінус 2», четвертого – «плюс vs мінус 3». Але цими типами потенціал засобів реалізації оцінного значення не вичерпаний. На морфологічному рівні він представлений словотвірними засобами: суфіксами демінутивності-пестливості із позитивнооцінним значенням (*ангел-**ятк**-о, зозул-**еньк**-а, красунь-**чик***) та аугментативності з пейоративним негативнооцінним суб'єктивним позаконтекстним значенням (*збор-**ищ**-е, звір-**юк**-а, злид-**от**-а*). Враховуючи периферійність і суб'єктивність цієї оцінки, лексемам даної групи, які вилучаються із словника публіцистики за словотвірною міткою, приписується оцінка «плюс vs мінус1». Те саме стосується префіксів **анти-, архі-, без-, не-** (***анти**-герой*, ***архі**-шахрай, **без**-грамотність, **не**-культурність*). Правда, деякі з оцінних префіксів є виразниками синкретичної (позитивної і негативної) суб'єктивної оцінки, її можна встановити лише в контексті вручну.

Правила сентимент-аналізу містять грунтовну лінгвістичну інформацію, яка застосовується після роботи морфолого-синтаксичного із зняттям омонімії і морфемно-словотвірного модуля комп'ютерної граматики української мови АГАТ [2].

Факторами, які впливають на характер сентименту, можна також вважати інтенсифікатори – прислівники (*дуже, мало, ледве* тощо), модифікатори (напр., заперечна частка *не)*, які міняють оцінку на протилежну. Ці фактори становлять ще одну групу правил, які враховані програмно на підставі синтаксичного аналізу тексту із проставляннням синтаксичних зв'язків в аналізованому тексті. Їм присвоюється оцінка –1, яка коригується контекстуально і вручну.

До сентимент-словника ми включали не тільки слова, які мають у денотативному значенні оцінку, а й слова, які асоціюються з чимось хорошим або поганим, тобто мають оціночні конотації (*черга, корок, безробіття*). Для автоматичного виявлення слів, які мають негативні або позитивні конотації, використовуються колокації (*боротися з, боротися за* тощо) – стійкі сполуки слів, які досить часто зустрічаються як з ідіоматичним значенням, так і з неідіоматичним, але несуть тональність чи посилюють її, а також оцінні фразеологізми. Оскільки сполуки співвідносяться із класами слів з оцінною семантикою і виражають позитивну або негативну оцінку, їм приписувалася лексема, що передає значення сполуки і, відповідно, сентимент: напр., *бити байдики* – ***лінуватися*** (негативна оцінка); *з глузду з'їхати* – **збожеволіти (**негативна).

Вище відзначалося, що при визначенні тональності тексту неабияке значення має об'єкт оповідання – про що йде мова у тексті? Сентимент-аналіз проводиться на основі ключових слів: якщо в тексті йдеться про газ, пенсії, зарплату або вугілля тощо, то словосполучення *підвищити ціну, зарплата зросте, збільшення пенсій, немає дефіциту вугілля, бунтують військові пенсіонери* включаються до словника, хоча кожне з цих ключових слів не входить до сентиментних слів і не містить позитиву саме по собі. З іншого боку, якщо вони будуть зустрічатися із словами, які позначатимуть позитивні якості (*максимальний, високий, великий* тощо), це додаватиме позитивну тональність (*зарплата висока, максимальна, велика*), і навпаки, зі словами *втрата ризик, черга* і под., якщо зустрічатимуться з прикметником *мінімальний* і под. будуть мати позитивну тональність. Формулюються сентиментні правила: сила сентименту в таких випадках коливається від +1 до -1 для всієї групи слів: *підвищити ціну* (-1)*, зарплата зросте* (+1)*, збільшення пенсій* (+1)*, немає дефіциту вугілля* (+1 , тому що передує предикатив із заперечною часткою ***не***)*, бунтують військові пенсіонери* (-1) тощо. Поточна версія алгоритму не враховує силу сентименту і має однакове значення (+1 versus -1). Ключові слова разом із сентимент-контекстом, тобто синтаксично зв'язаними словами, зберігаються у відповідній таблиці №2 (поки що їх більше 200). Якщо слово із контекстом цього списку зустрілося у тексті, йому приписується відповідне значення тональності (напр., *зростання тарифів (цін/и), немає сенсу (можливості, коштів), сумнівно етичний, агенти Кремля, політичні торги* тощо) яким приписується негативна тональність -1. Вони вибиралися в автоматизованому режимі з інформаційних новинних текстів, адресованих максимальній аудиторії телебачення. Зупинимося спеціально й на інверсії сентименту: якщо слово має позитивну тональність, то заперечення модифікує його на негатив і навпаки – при запереченні негативної тональності слово/сполучення отримує нульову тональність сентимента. Але ще треба спеціально укласти списки слів позитивних із запереченням і слів негативних із запереченням.

Виявлена вручну тональна лексика разом з контекстом вживання, приписаною їй оцінкою (позитивною vs негативною), джерелом тональності (експресія, емоція, факт) складає список №2, який зберігається в ексесівській таблиці і разом із словником оцінних слів є лінгвістичними засадничими матеріалами сентимент-аналізу.

Тепер перед нами стоїть завдання на текстових вибірках перевірити алгоритм сентимент-аналізу, бази даних, шкалу тонування, зручність підтримки системи, під якою розуміємо зручність коригування на текстах нових тематик. Система налаштована на теми політичних та економічних новин стрічок, що транслюються телебаченням, статей у газетах різних регіонів України. Значення сентименту речення **Is** у системі обчислюється як середньоарифметичне значень сентиментів слів, які до нього входять зі знаком плюс або мінус. Відповідно значення сентименту цілого тексту **It** дорівнює середньоарифметичному значенню сентименту речень, для яких обчислено тональність.

$$I_s = \frac{\Sigma_i^n |K| \times f_i}{\Sigma n_i}$$, де

Is – індекс тональності речення;

K – коефіцієнт тональності (від +3 до -3);

f – абсолютна частота тональності слів;

n – кількість слів у реченні.

$$I_t = \frac{\Sigma |I_S|}{\Sigma m_i}$$, де

I_t - індекс тональності тексту;

m_i- кількість слів у тексті.

Позитивний або негативний сентимент рахується окремо і подаються індекси позитивної або негативної тональності. Коливання значень від 0 до +1 або -1. Оскільки індекси – це відносна частота, множимо її на 100% і одержуємо значення у відсотках.

Отже, встановлений потенціал засобів реалізації сентименту в українських мас-медіа дозволить, по-перше, сформувати статистику найчастотніших моделей вираження сентименту; по-друге, охарактеризувати різні тематики з точки зору властивих для них лексичних категорій; по-третє, використовувати кількість позитивних, негативних і нейтральних слів із врахуванням шкалування для визначення границі між сентиментним і нейтральним текстом, а також для диференціювання досить значимих негативних і позитивних сторін описуваних сутностей.

ЛІТЕРАТУРА

1. Баранов А.Н. Введение в прикладную лингвистику. М. 2003. С. 213
2. Дарчук Н.П. Комп'ютерне анотування українського тексту: результати і перспективи / Дарчук Н. П. — К. : Освіта України, 2013. — 543 с.
3. Николаев И.С., Митренина О.В., Ландо Т.М. (ред.) Прикладная и компьютерная лингвистика. М.: Ленанд, 2016. — 316 с.
4. Онищенко І.В. Категорія оцінки та засоби її вираження в публіцистичних та інформаційних текстах. : автореф. дис. ... канд. філол. наук / Онищенко Ірина Володимирівна; Дніпропетр. націон. ун-т, Дніпропетровськ, 2004. — 22 с.
5. Петрик В. М. Сугестивні технології маніпулятивного впливу / [В. М. Петрик, М. М. Присяжнюк, Л. Ф. Компанцева та ін.]. – К. : Науково-видавничий відділ Національної академії СБ України, 2010. – 248 с.
6. Система Сентиментного анализа ATEX, основанная на правилах, при обработке текстов различных тематик Паничева П. В. (ppolin86@gmail.com) EPAM Systems, Санкт-Петербург, Россия
7. Янина А. О., Воронцов К. В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге // Машинное обучение и анализ данных.— 2016.— Т. 2, № 2.— С. 173–186.

REFERENCES

1. Baranov A.N. Vvedeniye v prikladnuyu lingvistiku. M.2003. C 213
2. Darchuk N.P. Kompiyuterne anotuvannya ukrayinskoho tekstu: resultaty i perspektyvy / Darchuk N.P. — K. : Osvita Ukrayiny, 2013/ — 543 s.
3. Nikolayev I.S., Mitrenina O.V., Lando T.M. (red.) Prikladnaya i kompiyuternaya lingvistika. M.: Lenand, 2016. — 316 s.
4. Onishchenko I.V. Kategoriya otsinky ta zasoby yiyi vyrazhennya v publitsystychnykh ta informatsiynykh tekstakh.: avtoref. dis. … kand. filol. nauk / Onishchenko Iryna Volodymyrivna; Dnipropetr. nation. un-t, Dnipropetrovs'k , 2004 — 22 s.
5. Petrik V.M. Sugestyvni tekhnologiyi manipulyatyvnoho vplyvu / [V. M. Petryk, M. M. Prysyazhnyuk, L. F. Kompantsevata in.]. – K. : Naukovo-vydavnytchiy viddil Natsionalnoyi akademii SB Ukrayiny, 2010. – 248 s.
6. Systema Sentimentnogo analiza ATEX, osnovannaya na pravilakh, pri obrabotke tekstov razlitchnykh tematik.Panicheva P.V. (ppolin86@gmail.com) EPAM Systems, Sankt-Peterburg, Rossiya.
7. Yanina A. O., Vorontsov K. V. Multomodalnyie tematitcheskiye modeli dlya razvedotchnogo poiska v kollektivnom bloge // Mashynnoye obutcheniye I analiz dannykh. 2016.— T. 2, № 2.— S. 173–186.

Linguistic approach for development of computer-based sentiment analysis in the Ukrainian language

N. Darchuk

Abstract. The article presents linguistic approach for development of computer-based sentiment analysis of collection of news, related commentary and feature materials from the Ukrainian social media. The sentiment is an emotional way of representing of information that can influence conscience of a reader or intentionally capture the mind of an addressee. Alongside, the sentiment analysis is aimed to emphasize not what is said, in particular or core information, fact or event, but how emotionally it is presented. There was determined a potential of use cases and means of sentiment implementation in the Ukrainian language.

Keywords: *sentiment analysis, sentiment object, sentiment subject, sentiment aspect, sentiment index of a sentence, sentiment index of a text.*