

# Методи розпізнавання сарказму в тексті

# Аналіз тональності

# Аналіз тональності тексту

**Ана́ліз тона́льності те́ксту** (*Sentiment analysis*) — клас методів контент-аналізу в комп'ютерній лінгвістиці, призначений для автоматизованого виявлення в текстах емоційно забарвленої лексики і емоційної оцінки авторів (думок) по відношенню до об'єктів, мова про які йде в тексті.

Тональність — емоційне ставлення автора висловлювання до деякого об'єкту (об'єкту реального світу, події, процесу або їх властивостями / атрибутам), виражене в тексті. Емоційна складова, виражена на рівні лексеми або комунікативного фрагмента, називається лексичною тональністю (або лексичним сентиментом). Тональність всього тексту в цілому можна визначити як функцію (в найпростішому випадку суму) лексичних тональностей складових його одиниць і правил їх поєднання.



# Аналіз тональності

аналіз тональності тексту зазвичай включає в себе наступні основні завдання:

- визначення наявності емоційного забарвлення;
- визначення полярності тексту;
- вилучення аспектів з емоційно забарвленого тексту



# Суб'єктивність та емоції

Існують дві концепції, що є дуже тісно пов'язаними з класифікацією емоційного забарвлення думок – суб'єктивність та емоції.

Об'єктивне речення визначає певну фактичну інформацію щодо навколишнього світу, в той час як суб'єктивне твердження висловлює почуття та думки окремої людини.

Прикладом об'єктивного твердження є наступне речення: «iPhone є продуктом компанії Apple».

Прикладом суб'єктивного твердження є наступне речення: «Мені 18 подобається iPhone»

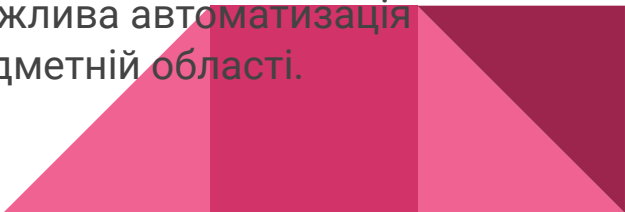


# Метод, оснований на словниках

Кожному слову чи фразі привласнена полярність та сила, і включає в себе інтенсифікацію і заперечення, щоб обчислити емоційне забарвлення кожного документу. Найбільшою проблемою методів, заснованих на словниках і правилах, є важкість процесу складання словника.

Для одержання методу, що класифікує документ з високою точністю, терміни словника повинні мати вірну вагу, адекватну предметній області документа. Наприклад, слово «непередбачуваний» по відношенню до сюжету фільму є позитивною характеристикою, але негативною по відношенню до, наприклад, політика.

Тому даний метод вимагає значних затрат часу людини, через те, що для хорошої роботи системи необхідно скласти велику кількість правил. Часто можлива автоматизація складення словників, проте зазвичай лише у дуже вузькій предметній області.



Слово	Тональність (1–9)
happy	8.11
good	7.37
dull	2.85
angry	2.75
sad	1.51

	Precision	Recall	Accuracy
Словниковий метод	0.59	0.685	0.681
НБК	0.643	0.771	0.772
SVM	0.65	0.769	0.77
НБК з використанням біграм	0.683	0.782	0.791
SVM з використанням біграм	0.684	0.778	0.793
Метод Хе Юлан та Жоу Деу (з класифікатором НБК)	0.653	0.78	0.781




# Сарказм




# SpaCy

SpaCy — одна з найбільш популярних опенсорсних НЛП-бібліотек. Написана на Cython, вона дуже добре оптимізована і призначена для використання в реальних проектах. Розглянемо на прикладах деяких з її можливостей.

Токенізація тексту — простими словами, бібліотека дозволяє розбити текст на смислові сегменти: слова, статті, пунктуацію. Позже ці сегменти — разом або окремо — можна представити у вигляді векторів для подальшого порівняння. В якості простого прикладу слова «кіт» і «пухнастий» у багатомірному векторному просторі окажуться ближче, чим той же «кіт» і «космічний корабель».



# Складніші випадки

- Я не не люблю старі катери з каютами. (Уловлювання заперечення)
  - Мені не подобається керування судном. (Заперечення, перевернутий порядок слів)
  - Я би дійсно дуже хотів би піти прогулятись у таку погоду! (Можливий сарказм)
  - Кріс Крафт виглядає краще, ніж Лаймстоун (Дві торгові марки, що роблять визначення цілі дуже важким)
  - Кріс Крафт виглядає краще, ніж Лаймстоун, але Лаймстоун розробляє мореплавність та надійність. (Дві торгові марки, дві позиції)
- 

Під час живої розмови сарказм можна легко визначити за допомогою міміки, жестів і тону оратора.

Виявлення сарказму в текстовому спілкуванні — незвичне завдання, оскільки жоден із цих сигналів не є доступним



# Datasets

## Riloff dataset

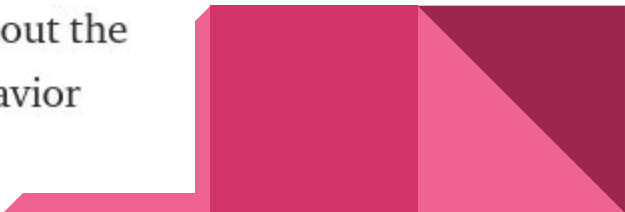
Consists of 3.2k tweet IDs, manually labelled by third party annotators. The labels capture the subjective perception of the annotators (perceived sarcasm). We collected tweets using the Twitter API as well as the historical timeline tweets for each user. For a user with tweet  $t$  in Riloff, we collected those historical tweets posted before  $t$ . Only 701 original tweets along with the corresponding timelines could be retrieved.

## Ptacek dataset

Consists of 50k tweet IDs, labelled via distant supervision. Tags used as markers of sarcasm are #sarcasm, #sarcastic, #satire and #irony. This dataset reflects intended sarcasm. In a similar setting as with Riloff we could only collect 27.1k tweets and corresponding timelines.



# Steps

1. Cleaning: removing other hashtags, links, images.
  2. Lexical features: removing stopwords, tokenizing, BoW, POS tagging
  3. Pragmatic features: smileys/emojis, mentions
  4. Context incongruity: the context not being in agreement. Can be explicit and implicit
  5. User embeddings: stylometric and personality features. This embeddings encode a Twitter users' information in a way that two similar users have a similar embedding. Because they tweet about the same topic, write with a similar style, or have comparable behavior patterns.
- 

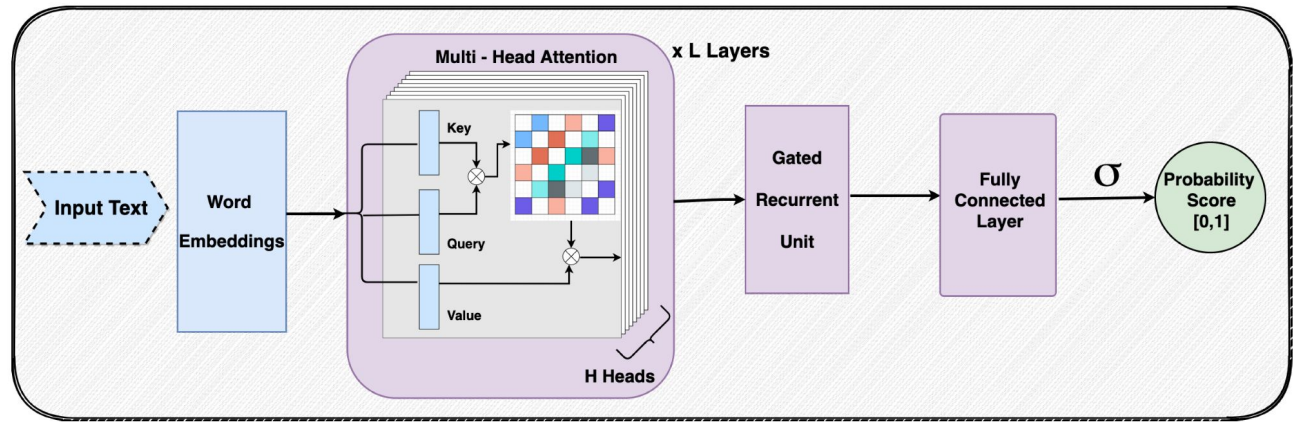
## Exploring Author Context for Detecting Intended vs Perceived Sarcasm, Oprea and Magdy, 2019

- Чи передбачає поведінка користувача саркастичний характер твіту?
- Яку роль відіграє embedding користувачів у наборах даних Riloff проти Ptacek?
- Яка різниця у результаті між двома наборами даних?

Чи можемо ми передбачити сарказм, просто знаючи минулу поведінку та особистість користувача, не дивлячись на твіт. Чи є сарказм випадковим чи є певні особливості особистості, які роблять когось більш імовірним саркастичним?

І якщо так, то який їх вплив на два набори даних? Чи допомагає нам пізнання когось, його хобі та інтересів, його характеру виявити, чи є їх твіти більш саркастичними?

Model	Riloff	#Riloff
EX-CASCADE	0.457	0.818
EX-W-CASCADE	0.478	0.797
EX-ED	<b>0.545</b>	<b>0.827</b>
EX-SUMMARY	0.492	0.772



The team implemented its model in PyTorch, a deep-learning framework in Python (Sarcasm detection with Python).

The approach it took to building the models was based on five components: data pre-processing, multi-head self-attention module, gated recurrent unit module, classification, and model interpretability. What's more, the experiments were conducted on multiple datasets from varied data sources and show significant improvement over the state-of-the-art models by all evaluation metrics.