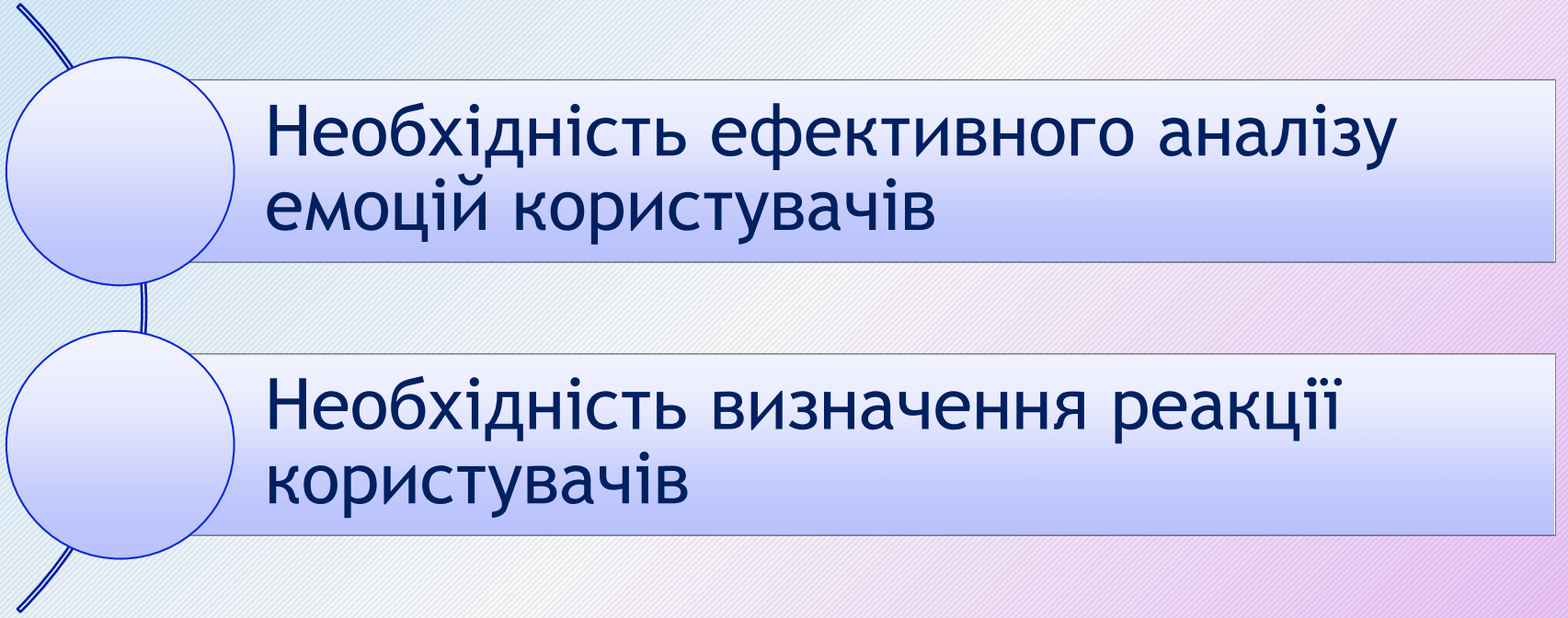
Оцінювання емоційного відклику на новини за коментарями

Олашин О. О. КА-21
Шипік Д. В. КА-21
Дипломний керівник: Дідковська М. В.

Актуальність роботи



Об’єкт та предмет дослідження

Об’єкт: підходи та методи аналізу тональності.

Предмет: системи оцінювання емоційного відклику.

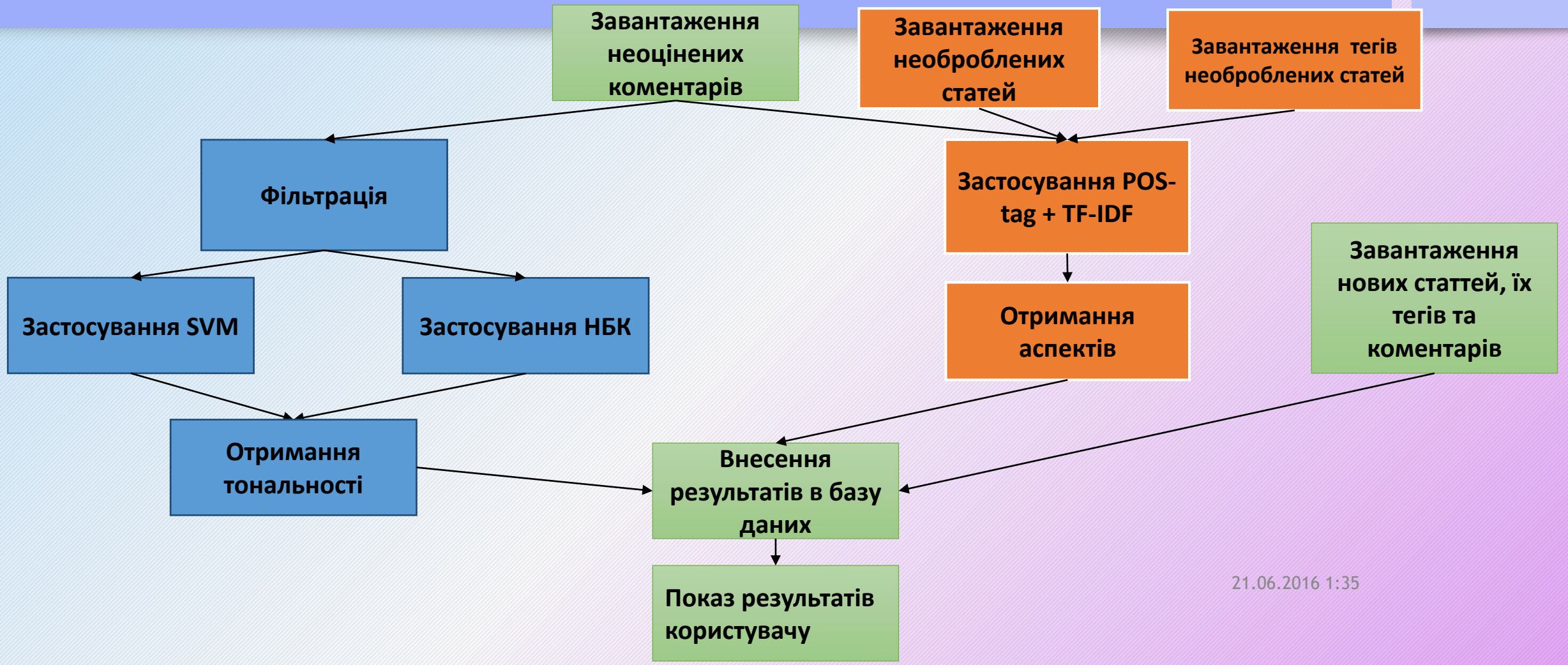
Мета роботи

- Створити веб-додаток в якому користувачі в режимі онлайн можуть проводити пошук новин по ключовим словам та отримувати оцінку емоційної реакції інших користувачів.

Постановка задачі дипломної роботи

- Проаналізувати існуючі методи аналізу тональності тексту: методи емоційної оцінки та визначення аспетків.
- На основі аналізу запропонувати власні алгоритми аналізу тональності та алгоритми визначення аспектів.
- Проаналізувати існуючі реалізації роботи з оцінкою реакції на емоції.
- Розробити програмний продукт, що реалізує оцінювання емоційної реакції на новини за коментарями.

Структурна схема



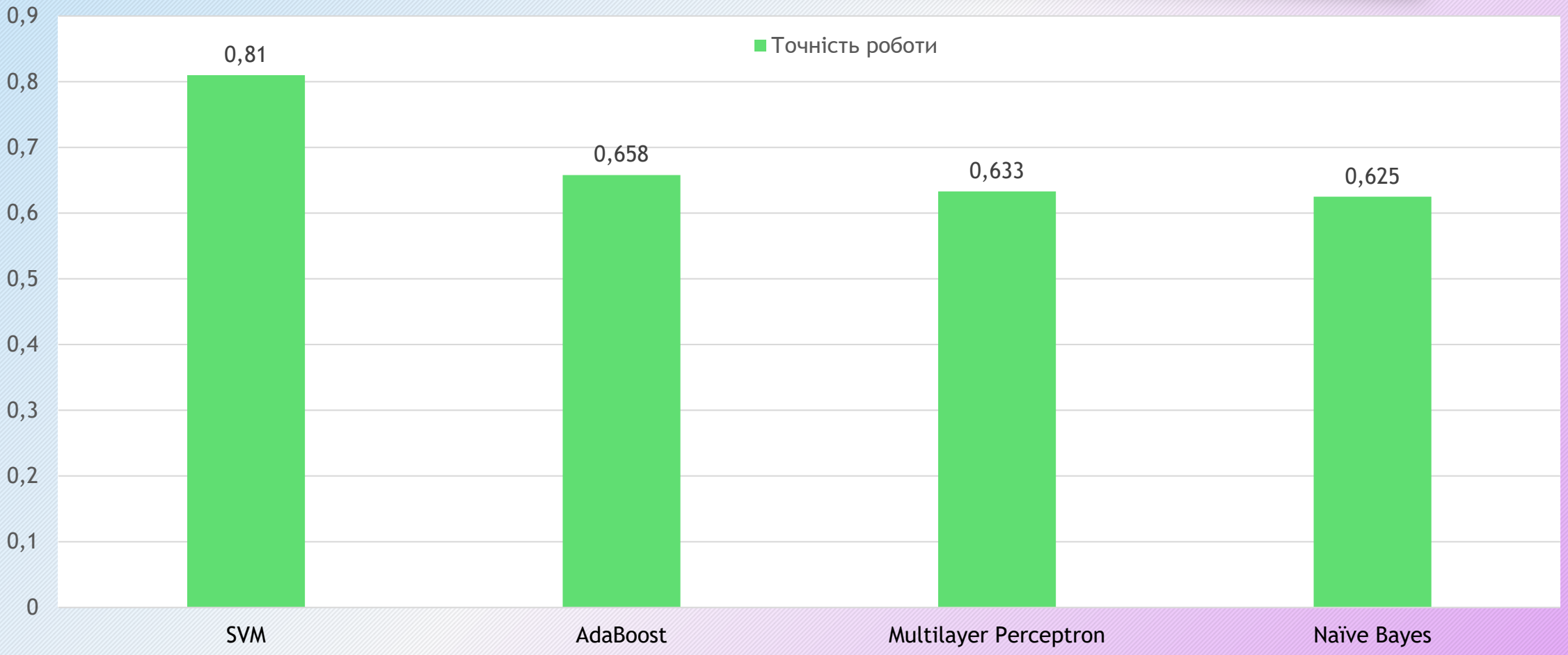
Аналіз існуючих підходів

- Машинне навчання з учителем
- Машинне навчання без вчителя
- Метод, заснований на теоретико-графових моделях

Недоліки

- Важкість складання тренувальних даних для навчання
- Важкість складання словника
- Складність реалізації та повільна швидкість роботи

Порівняльний аналіз точності існуючих методів



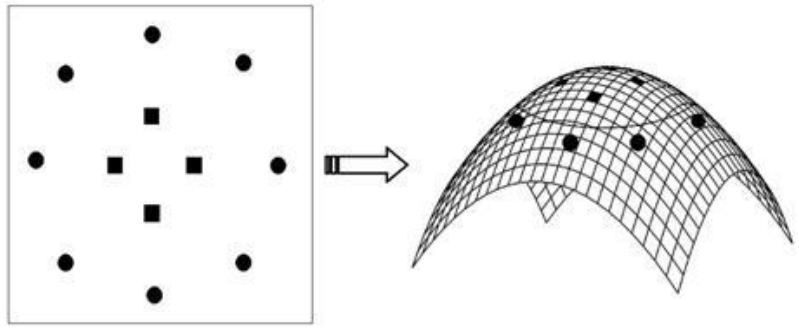
SVM

- Метод опорних векторів (SVM) – метод навчання з учителем, який аналізує та класифікує дані.

$$\sum_{i=1}^{N} \alpha_i d_i K(\mathbf{x}, \mathbf{x}_i) = 0.$$

$$\mathbf{w}_o = \sum_{i=1}^{N} \alpha_{o,i} d_i \varphi(\mathbf{x}_i),$$

$$K(\mathbf{x}, \mathbf{x}_i) = \varphi^T(\mathbf{x})\varphi(\mathbf{x}_i)$$



SVM

Переваги:

- метод зводиться до вирішення задачі квадратичного програмування у випуклій області, яка завжди має єдиний розв’язок;
- метод знаходить роздільну смугу максимальної ширини, що дозволяє надалі здійснювати кращу класифікацію

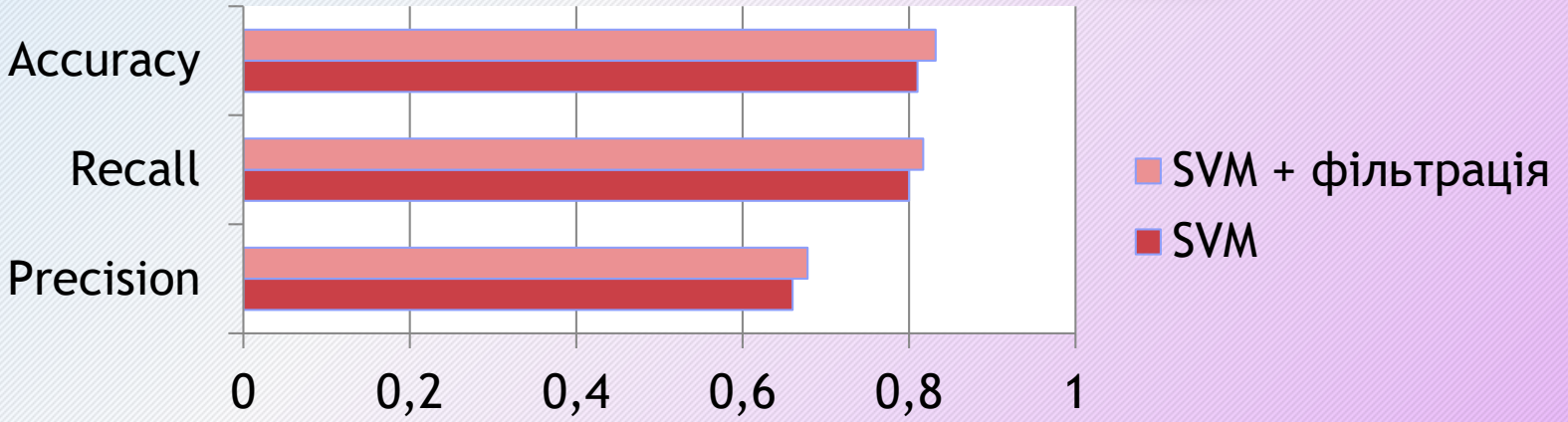
SVM

Недоліки:

- метод чутливий до шумів та стандартизації даних;
- повільне навчання.

Пропонується: SVM + фільтрація

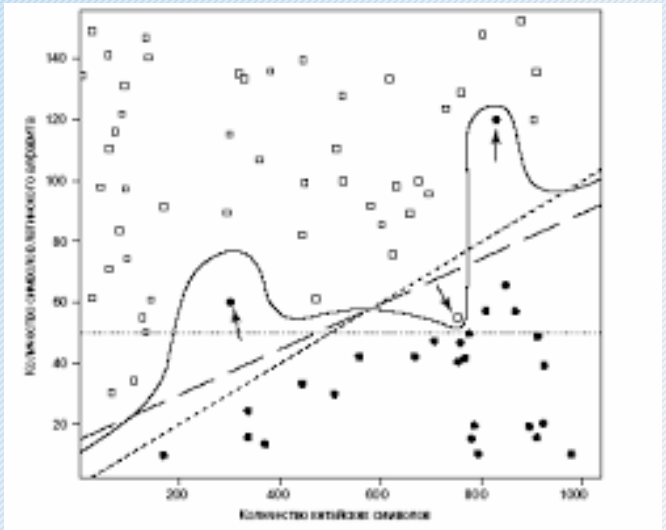
Використання фільтрації покращує роботу SVM на:
Precision – 2.7%,
Recall – 2.1% та
Accuracy – 2.7%



$$Precision = \left(\frac{TP}{TP + FN}\right) \; Recall = \left(\frac{TP}{TP + FP}\right) \; Accuracy = \left(\frac{TP + TN}{P + N}\right)$$

Швидкість роботи алгоритмів

Наївний байєсівський класифікатор



$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

$$p(C|F_1, \ldots, F_n) = \frac{p(C)\, p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)}.$$

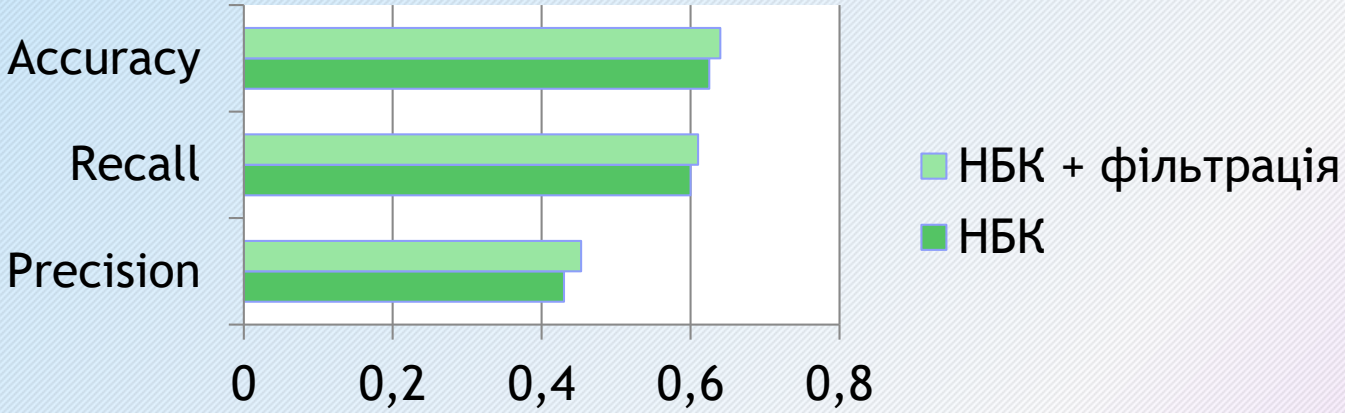
$$p(C_k|x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

Наївний байєсівський класифікатор

Переваги – стійкість до шумів; висока швидкодія.

Недолік — через припущення про незалежність, втрачається точність.

Пропонується: НБК + фільтрація



НБК на:
Precision – 5.3%,
Recall – 1.6% та
Accuracy – 2.4%
і при цьому не суттєво збільшує використання системних ресурсів.

$$Precision = \left(\frac{TP}{TP + FN}\right) \; Recall = \left(\frac{TP}{TP + FP}\right) \; Accuracy = \left(\frac{TP + TN}{P + N}\right)$$

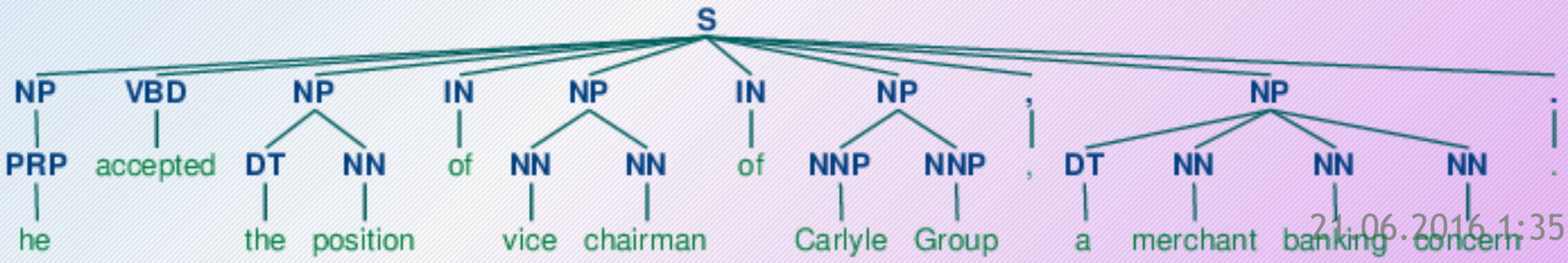
Фільтрація

- Перед відправкою даних в алгоритми НБК та SVM проводиться їх фільтрація:

1. Видалення стоп-слів (використання словника стоп-слів англійської мови).
2. А також наступні регулярні вирази:
   1) [a-zA-Z0-9\!\?\.\,\-\=\)\(]
   2) ([a-zA-Z])\1+([\!\?\.\,\-\=\)\(])\2+ => \1 \2
   3) ([a-zA-Z])\1+([0-9])\2+ => \1 \2
   4) ([0-9])\1+ ([\!\?\.\,\-\=\)\(])\2+ => \1 \2
   5) …
   6) ([0-9\.\,] {3,}) => ‘’
   7) ([a-zA-Z])\1{3,} => \1\1
   8) ([\?\.\,][\?\.\,])+ => \1\1
   9) (\!\?)[\?\!] => \1

POS-tag

- Цей метод знаходить явно виражені аспекти, які є іменниками та словосполученнями.
- Іменники та словосполучення (або групи) ідентифікуються з допомогою ідентифікатора частин мови (POS). Частота їх появи підраховується, і тільки найчастіші з них зберігаються.



POS-tag

Переваги: Легкий у застосуванні; Достатньо точний.

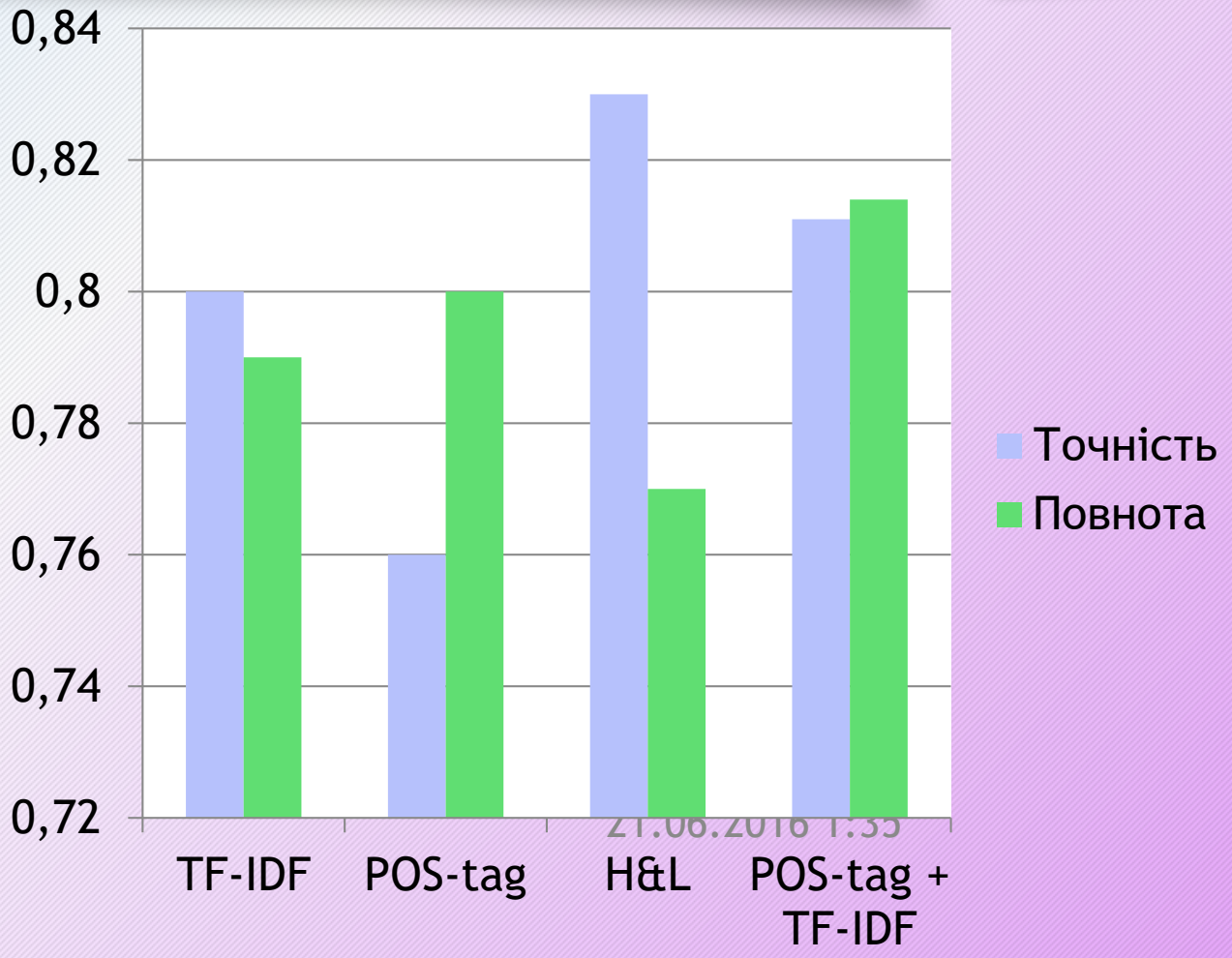
Недоліки: Пропускає не часті аспекти;
Сильно втрачає точність при визначенні аспектів, що складаються більше ніж з 2-х слів.

Пропонується: POS-tag + TF-IDF

- Комбінований варіант алгоритму полягає в наступному: спочатку відбувається визначення всіх іменників (pos-tag частина). Потім відбувається їх підрахунок за допомогою TF-IDF та відкидаються не релевантні іменники.
- Переваги: більша точність визначення аспектів та повнота визначення аспектів.
- Недоліки: збільшення часу роботи.

Результати роботи

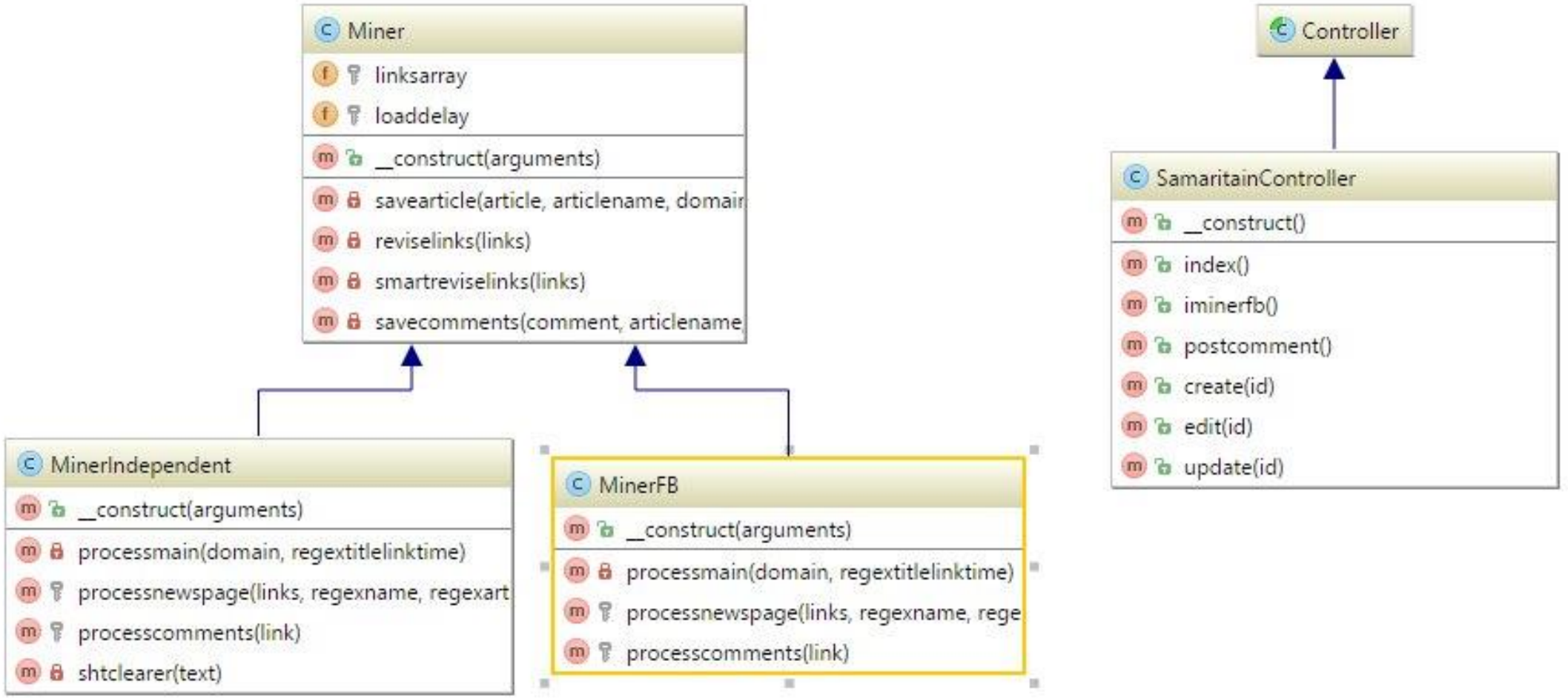
	Точність	Повнота
TF-IDF	0.8	0.79
POS-tag	0.76	0.8
H&L	0.83	0.77
POS-tag + TF-IDF	0.811	0.814



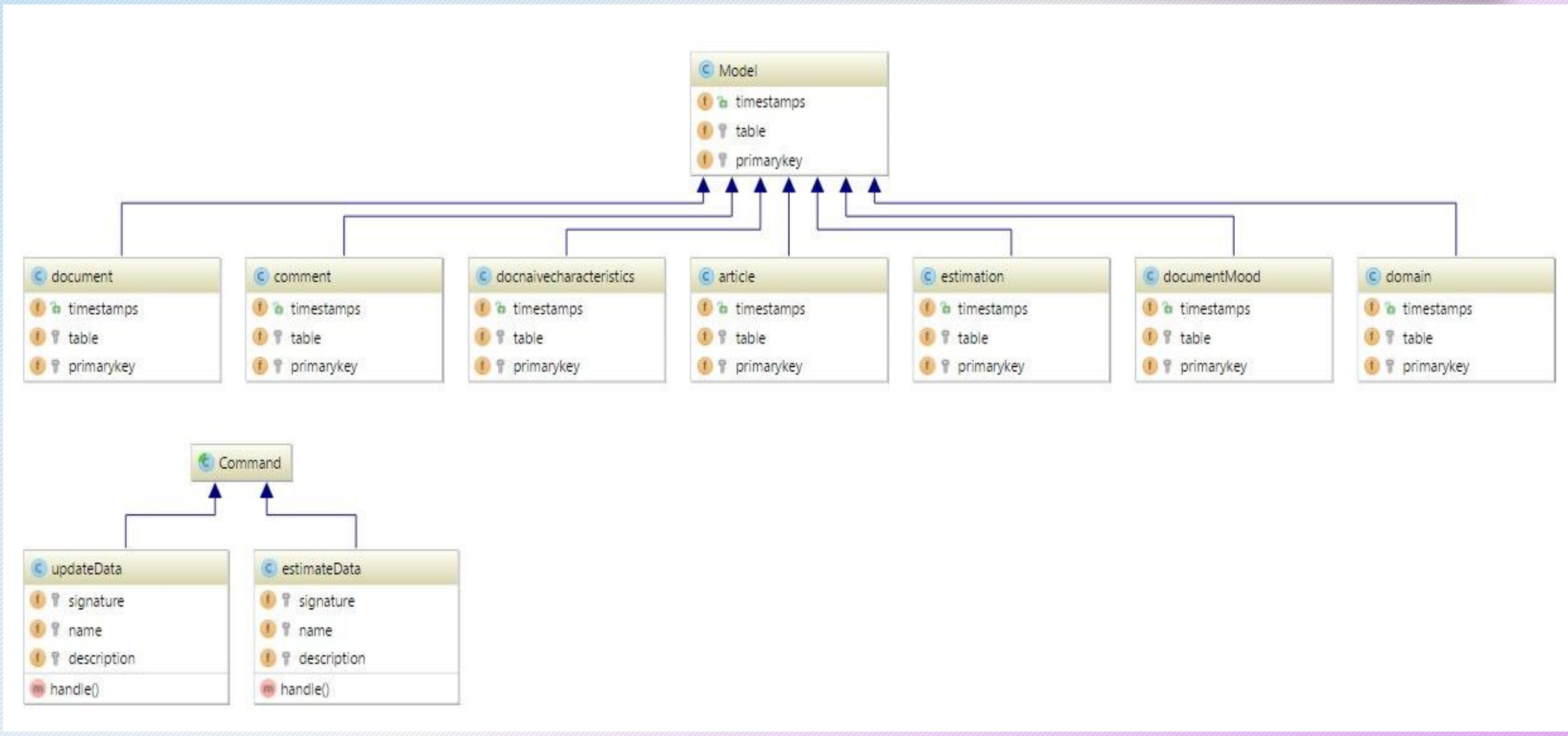
Вибір платформи та мови програмування

- В якості мов та інструментів були обрані php, python, laravel, nltk, scikit-learn.

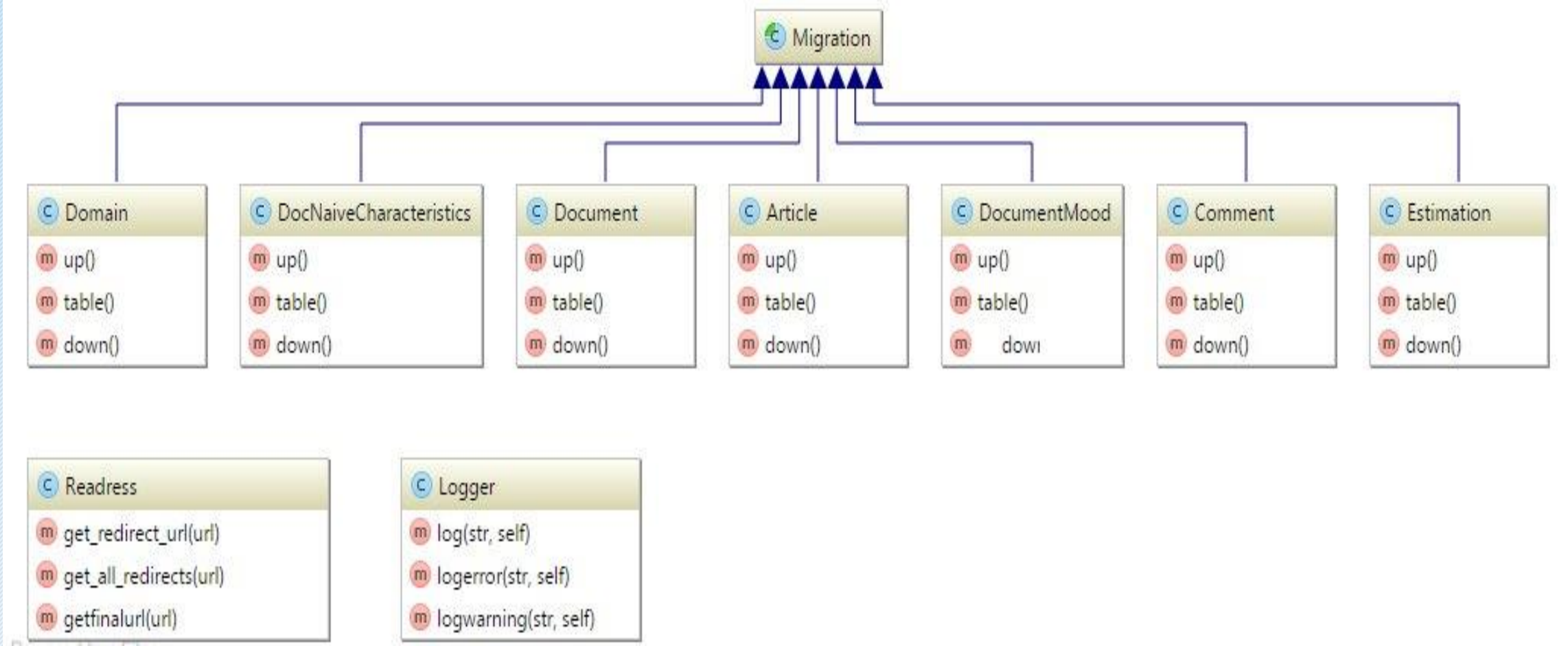
Архітектура програмного продукту



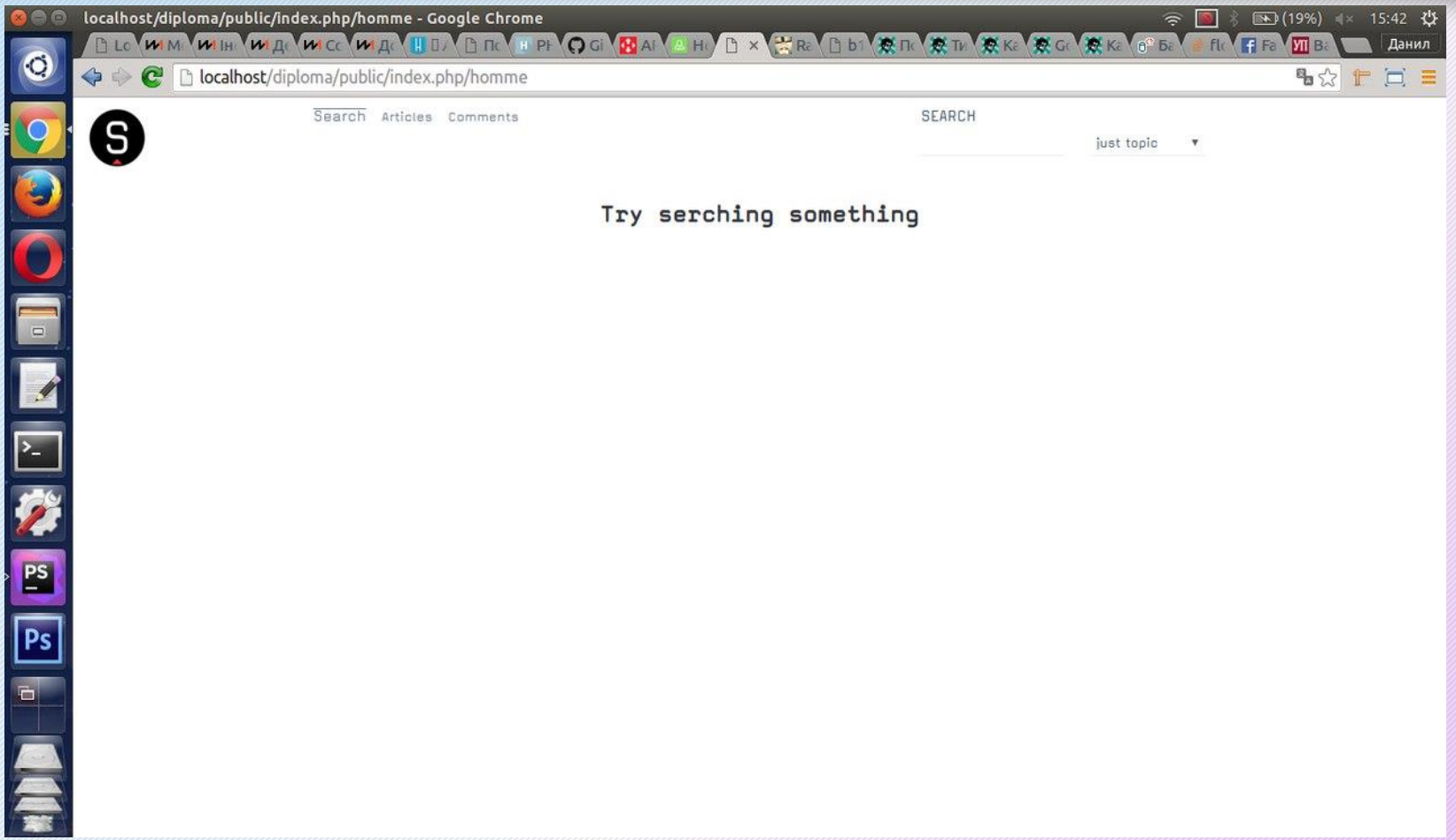
Архітектура програмного продукту



Архітектура програмного продукту

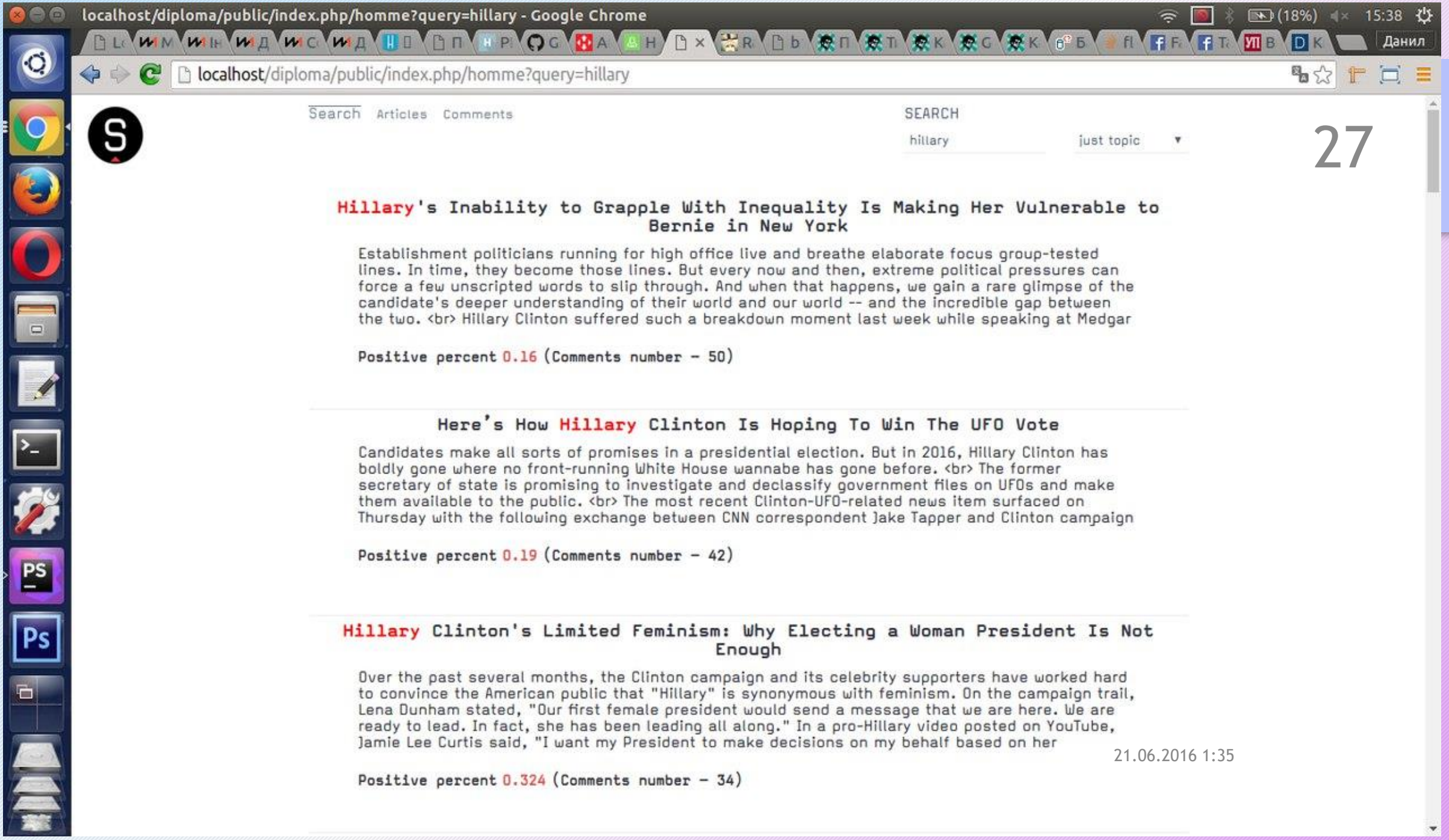


Інтерфейс програмного продукту



Веб-додаток для пошуку новин і розгляду всього спектру емоційної реакції на них.

Пошук новин відбувається по ключовим словам, що вводить користувач.


localhost/diploma/public/index.php/homme?query=hillary - Google Chrome
localhost/diploma/public/index.php/homme?query=hillary
Search Articles Comments
SEARCH
hillary
just topic
Hillary's Inability to Grapple With Inequality Is Making Her Vulnerable to Bernie in New York
Establishment politicians running for high office live and breathe elaborate focus group-tested lines. In time, they become those lines. But every now and then, extreme political pressures can force a few unscripted words to slip through. And when that happens, we gain a rare glimpse of the candidate's deeper understanding of their world and our world -- and the incredible gap between the two.
 Hillary Clinton suffered such a breakdown moment last week while speaking at Medgar
Positive percent 0.16 (Comments number - 50)
Here's How Hillary Clinton Is Hoping To Win The UFO Vote
Candidates make all sorts of promises in a presidential election. But in 2016, Hillary Clinton has boldly gone where no front-running White House wannabe has gone before.
 The former secretary of state is promising to investigate and declassify government files on UFOs and make them available to the public.
 The most recent Clinton-UFO-related news item surfaced on Thursday with the following exchange between CNN correspondent Jake Tapper and Clinton campaign
Positive percent 0.19 (Comments number - 42)
Hillary Clinton's Limited Feminism: Why Electing a Woman President Is Not Enough
Over the past several months, the Clinton campaign and its celebrity supporters have worked hard to convince the American public that "Hillary" is synonymous with feminism. On the campaign trail, Lena Dunham stated, "Our first female president would send a message that we are here. We are ready to lead. In fact, she has been leading all along." In a pro-Hillary video posted on YouTube, Jamie Lee Curtis said, "I want my President to make decisions on my behalf based on her
21.06.2016 1:35
Positive percent 0.324 (Comments number - 34)

localhost/diploma/public/index.php/comment?id=121 - Google Chrome

localhost/diploma/public/index.php/comment?id=121

Search Articles Comments

Comment

For article: Hillary's Inability to Grapple With Inequality Is Making Her Vulnerable to Bernie in New York
Domain: huffingtonpost.com

Sanders is staying in it to win the nomination, but win or lose, also so that progressive delegates can rewrite the Democratic party platform.

localhost/diploma/public/index.php/comment?id=123 - Google Chrome

localhost/diploma/public/index.php/comment?id=123

Search Articles Comments

Comment

For article: Hillary's Inability to Grapple With Inequality Is Making Her Vulnerable to Bernie in New York
Domain: huffingtonpost.com

As a lawyer, Hillary learned early on how to achieve results she was determined (or paid) to achieve, truth be damned. What she also learned was that she couldn't be true to any cause that could potentially disenfranchise the wealthy elites running America. Her campaign against Bernie is a campaign against everyone yearning and struggling for a better future. Belief in anything better is pie-in-the-sky to someone who gave up long ago in order to advance her own self interests. But for people who never gave up, who never gave in. . . our time has come. Ridding the Democratic race of inequality's ultra-adept apologist, Hillary Clinton, will be a watershed moment in the war against it. Give it everything you have people of New York! If Bernie wins your state, it will be her Waterloo.

Висновки

1. Запропоновано систему у аналізу новин за емоційною реакції на них, що складається з модулю аналізу емоцій та модулю автоматичного визначення аспектів.
2. Запропоновано модифікацію методів НБК та SVM (за допомогою фільтрації) та модифікацію алгоритму POS-tag з використанням TF-IDF.
3. Враховано особливості предметної області шляхом використання тегів новинних статей як аспектів.

Висновки

4. Модифікація методу НБК та SVM (використання фільтрації) призвела до наступних результатів: використання фільтрації покращує роботу SVM на: Precision – 2.7%, Recall – 2.1% та Accuracy – 2.7%; НБК на: Precision – 5.3%, Recall – 1.6% та Accuracy – 2.4%
5. Модифікація методу POS-tag (комбінування з TF-IDF) призвела до наступних результатів: покращення результатів порівняно з TF-IDF: точність - 1.3%, повнота - 3%; порівняно з POS-tag: точність - 6.7%, повнота - 1.7%.
6. Розроблено веб-сайт для аналізу емоційного відклику користвачів за коментарями.

Подальші шляхи роботи

Подальші дослідження можна проводи у напрямку визначення, аналізу та врахування бот-коментарів; розширення доступних для аналізу сайтів; врахування синонімічних аспектів.

Апробація роботи

1. Олашин О., Шипік Д.. Порівняння точності алгоритмів аналізу тональності на прикладі твіттів/ О. Олашин, Д. Шипік // Міжнародний науковий журнал. – 2016 - №5.

Дякуємо за увагу!