Інструментальні засоби аналізу та обробки великих масивів даних

Виконала студентка

КПІ ім.Ігоря Сікорського

Групи ТР-71 мп

Петрова Т.О.

Керівник: Гусєва І.І

Мета та задачі дослідження

Метою розробки є створення програмного продукту для аналізу та обробки великих масивів даних

Задачі:

- дослідити існуючі методи аналізу та обробки даних;
- обґрунтувати необхідність використання технологій великих даних;
- проаналізувати специфіку аналізу тональності твітів користувачів;
- пристосувати методи аналізу та методи обробки до даної предметної області;
- розробити програмне забезпечення для аналізу та обробки тональності великих масивів твітів.

Аналіз тональності текстів

Наївний баєсів класифікатор

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

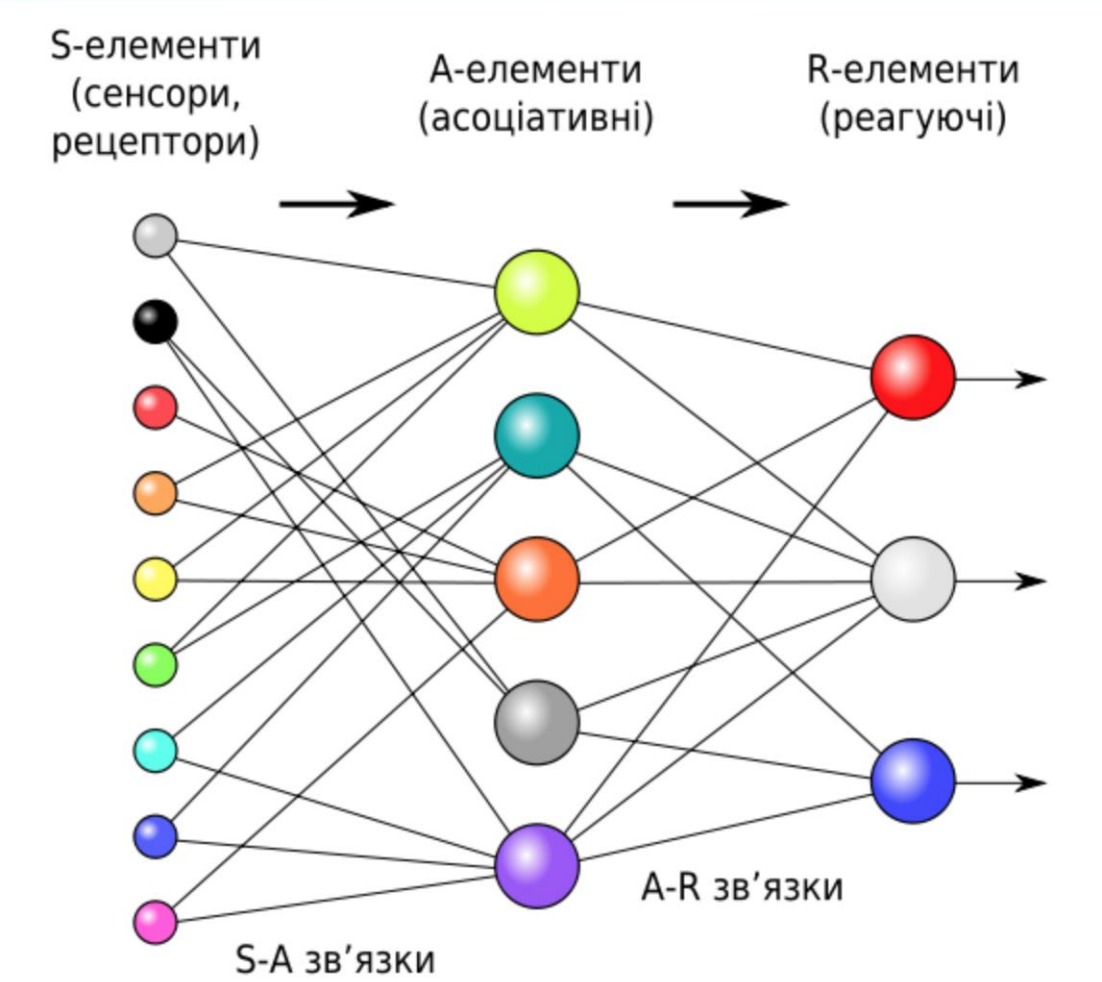
P(c|x) – апостеріорна ймовірність класу *c* при данному значенні ознаки *x*.

P (c) - апріорна ймовірність даного класу.

P (x | c) - правдоподібність, тобто ймовірність даного значення ознаки при даному класі.

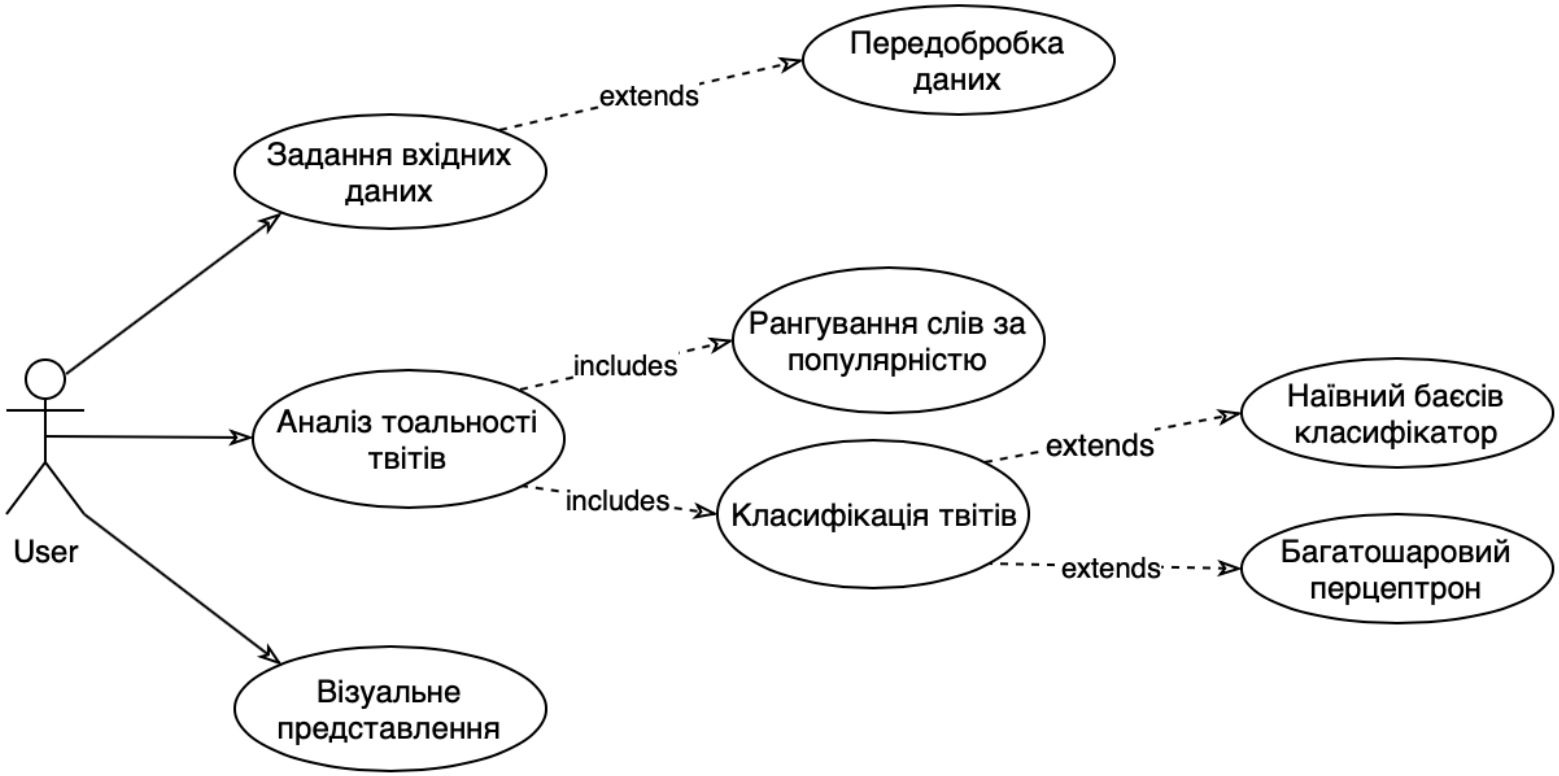
P (x) - апріорна ймовірність даного значення ознаки.

Багатошаровий перцептрон Румельхарта



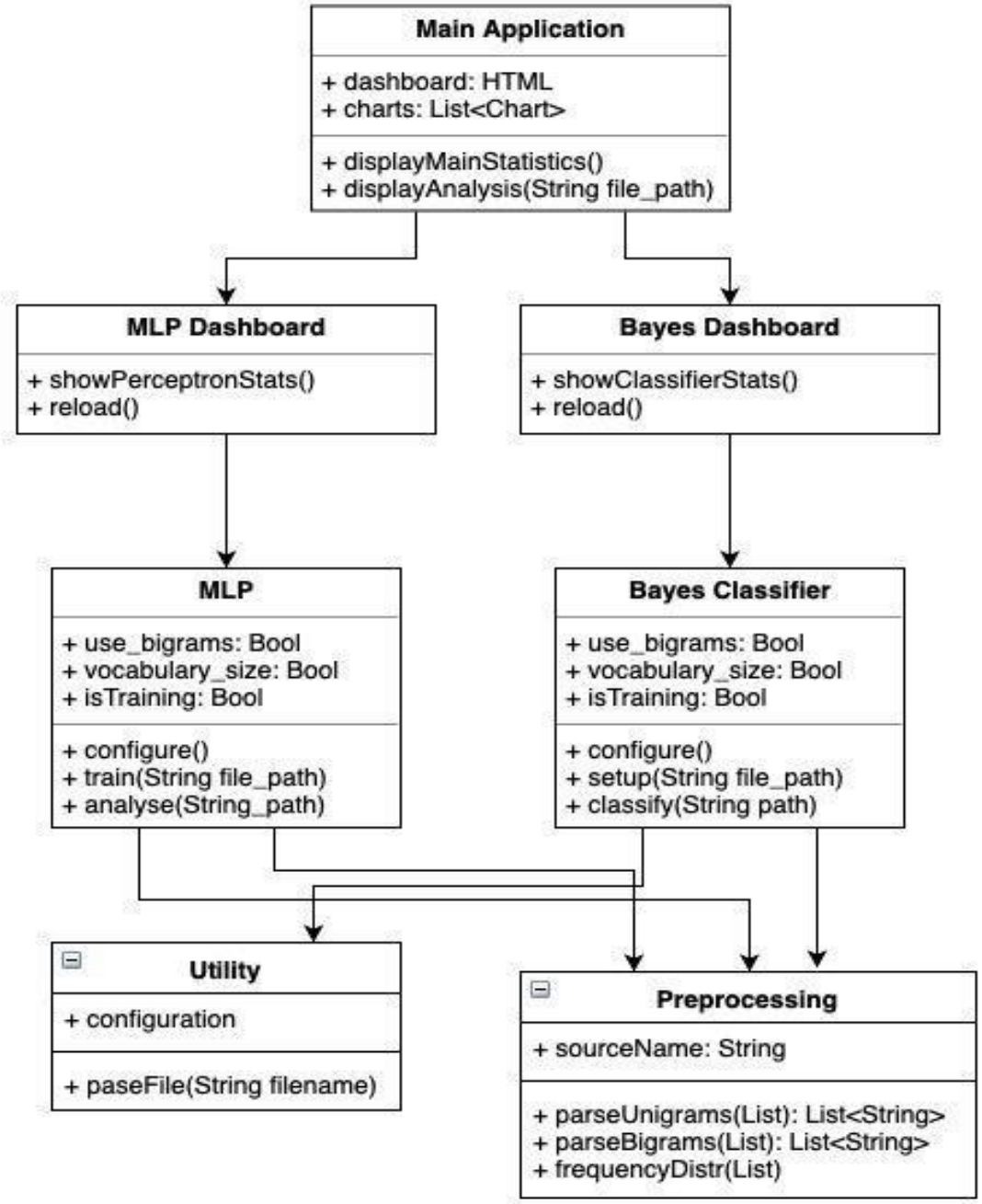
Функціональна декомпозиція системи

Діаграма прецедентів системи

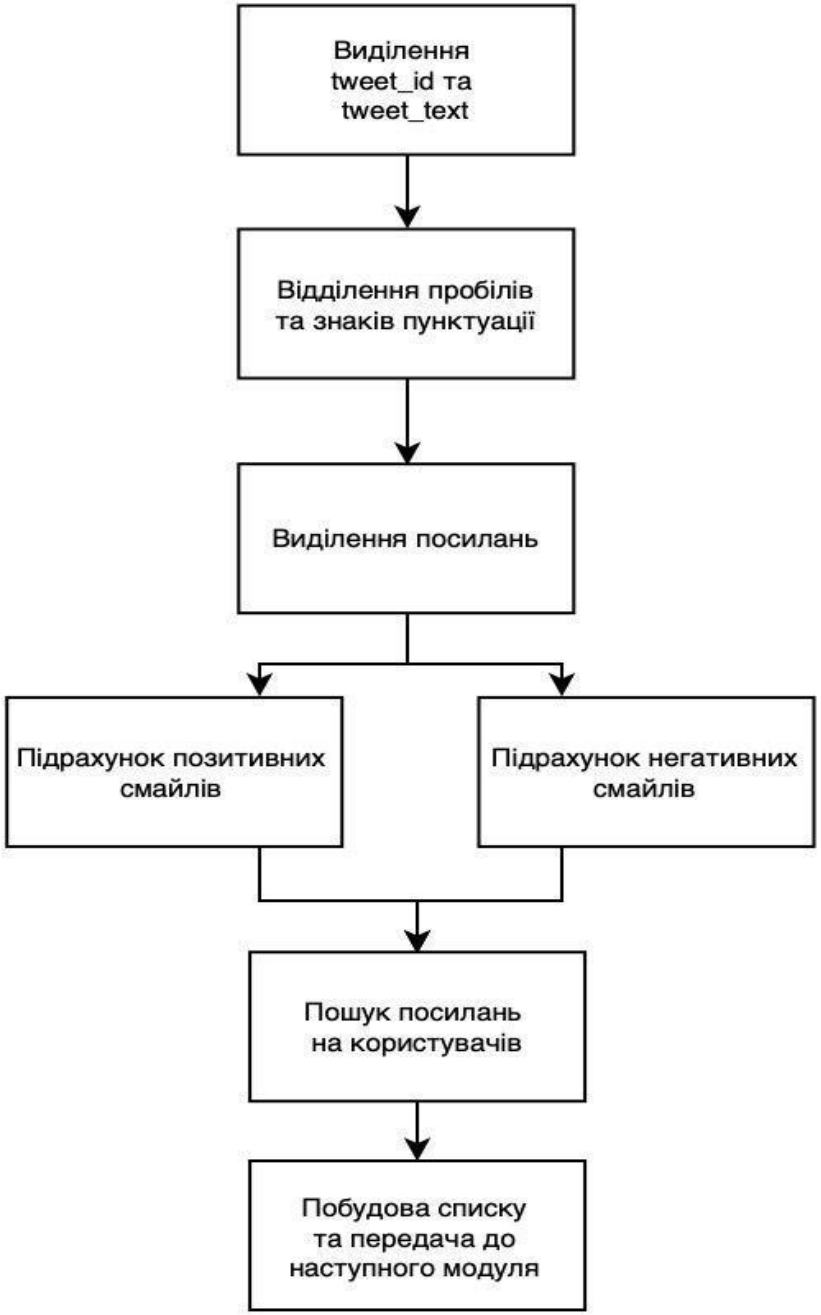


Архітектура системи

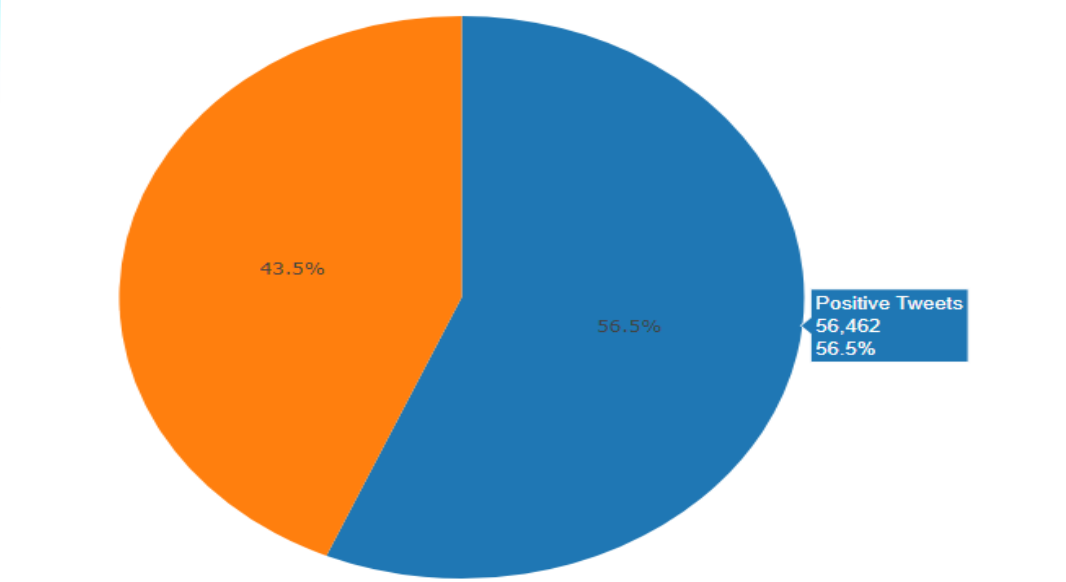
Діаграма класів

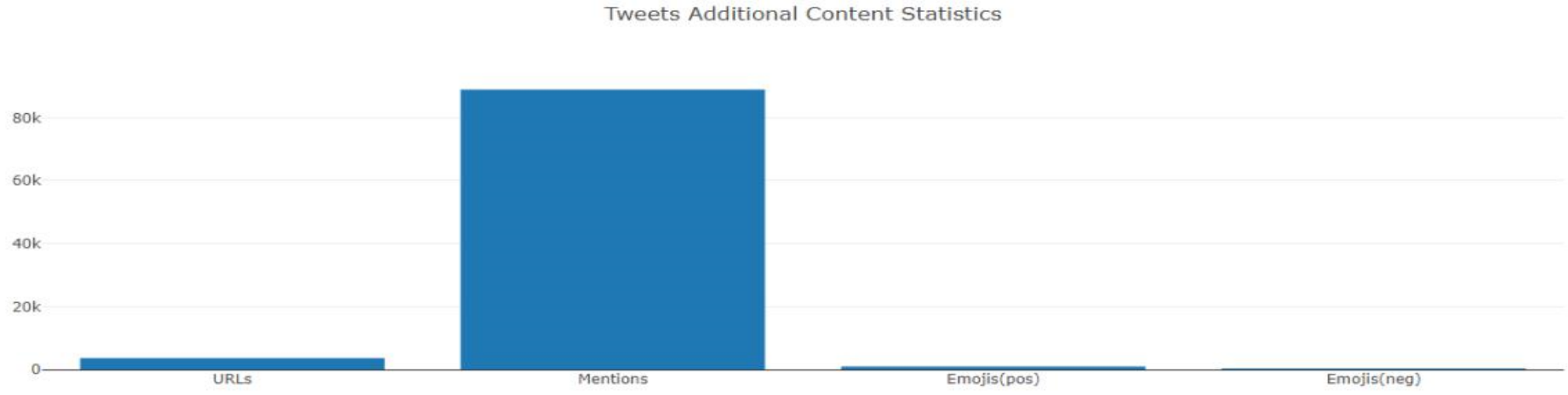


Алгоритм попередньої обробки набору даних



Приклади інтерфейсу користувача





Приклади інтерфейсу користувача

Most popular N-grams

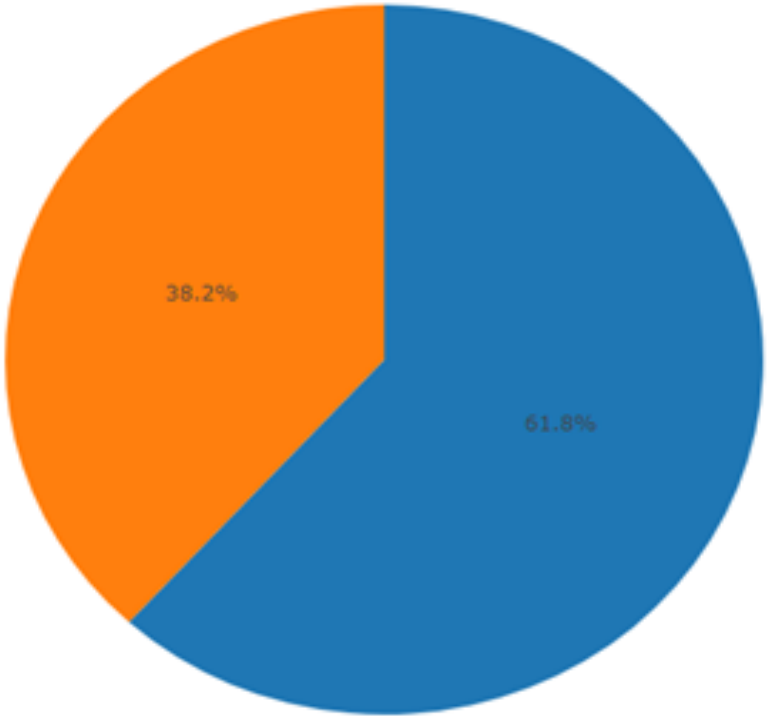
Top words
i
a
so
to
me
that
but
be
im
have
and
it
the
my
you
in
is
for
on
of

Top word pairs
have,to
for,the
i,love
have,a

Top tweets analysis

Tweet	Emotion
is so sad for my apl friend	0
i missed the new moon trailer	0
omg its already o	0
omgaga im soo im gunna cry ive been at this dentist since i was suposed just get a crown put on	0
i think mi bf is cheating on me t_t	0
or i just worry too much	0
juusst chillin	1
sunny again work tomorrow tv tonight	0
handed in my uniform today i miss you already	0
hmm i wonder how she my number USER_MENTION	1
i must think about positive	1
thanks to all the haters up in my face all day	1
this weekend has sucked so far	0
jb isnt showing in australia any more	0
ok thats it you win	1
this is the way i feel right now	0
awhhe man im completely useless now funny all i can do is twitter URL	0
feeling strangely fine now im gonna go listen to some	1

Приклади інтерфейсу користувача

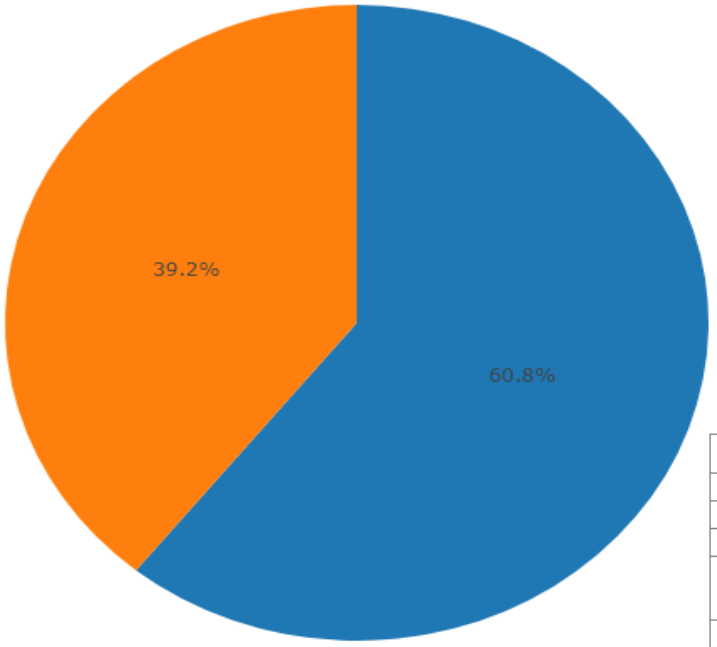


Тональність твітів за наївним баєсом

Top tweets analysis

Tweet	Emotion
huge roll of thunder just now so scary	0
i just cut my beard off its only been growing for well over a year im gonna start it over USER_MENTION is happy in the meantime	0
very sad about iran	0
wompp wompp	0
youre the only one who can see this cause no one else is following me this is for you because youre pretty awesome	1
level is i was writing a massive blog tweet on myspace and my comp shut down now its all lost in fetal	0
headed to hospitol had to pull out of the golf tourny in place i think i reripped something yeah that	0
boring EMO_NEG whats wrong with him please tell me	0
cant be bothered i wish i could spend the rest of my life just sat here and going to gigs seriously	0
feeling like shit right now i really want to sleep but noo i have hours of dancing and an art assignment to finish	0
goodbye exams hello alcohol tonight	0
i didnt realize it was that deep geez give a girl a warning atleast	0
i hate it when any athlete appears to tear an acl on live television	0
i miss you guys too i think im wearing skinny jeans a cute sweater and heels not really sure what are you doing today	1
meet your meat URL	1
my horsie is moving on saturday morning	0
no sat off need to work days a week	0

Приклади інтерфейсу користувача



Тональність твітів за перцептроном Румельхарта

Tweet	Emotion
is so sad for my apl friend	0
i missed the new moon trailer	0
omg its already o	0
omgaga im soo im gunna cry ive been at this dentist since i was suposed just get a crown put on	0
i think mi bf is cheating on me t_t	0
or i just worry too much	0
juusst chillin	1
sunny again work tomorrow tv tonight	0
handed in my uniform today i miss you already	0
hmm i wonder how she my number USER_MENTION	1
i must think about positive	1
thanks to all the haters up in my face all day	1
this weekend has sucked so far	0
jb isnt showing in australia any more	0
ok thats it you win	1
this is the way i feel right now	0
awhhe man im completely useless now funny all i can do is twitter URL	0

Порівняння результатів роботи

Алгоритм	Ознака наявності		Ознака частотин	
	Уніграми	Уніграми та біграми	Уніграми	Уніграми та біграми
Наївний баєсів класифікатор	78.16	79.68	77.52	79.38
Багатошаровий перцептрон	80.1	81.7	80.15	81.35

Висновки

- Метою роботи було створення та дослідження системи аналізу та обробки великих масивів даних. Було взято до уваги, що аналіз тексту є важливою сферою використання технології великих даних.
- У ході роботи досліджено методи аналізу та обробки даних, обрано два алгоритми для класифікації тексту: наївний баєсів класифікатор та багатошаровий перцептрон Румельхарта.
- На основі запропонованих алгоритмів аналізу було спроектовано та розроблено програмне забезпечення для визначення тональності тексту. Завдяки зіставленню словника найбільш популярних уніграм та біграм з тренувального набору та вхідних даних система визначає емоційне наповнення тексту.
- Алгоритм аналізу та обробки великих масивів даних розбито на декілька фаз роботи. Початкові етапи: зчитування вхідного файлу та передобробка забезпечують наступні етапи відформатованими, готовими до аналізу даними. Побудова словника та, безпосередньо, алгоритми аналізу виділено в окремі фази.
- Під час розробки системи було використано мову програмування Python та інтегроване середовище розробки PyCharm.
- Створена система відповідає поставленим задачам. Результати роботи впроваджено в науковій роботі кафедри (тема: № 0117U003798).