Розподілене обчислення і прогнозування важливості брендів в Інтернеті

Автор: студентка 6-го курсу

Групи КА-43м

Климковська Катерина

Керівник:

к.т.н. Діденко Д.Г.

Мета роботи

- Метою даного дослідження є створення алгоритму автоматичного розпізнавання тональності відгуків про бренд та прогнозування зміни важливості цього бренду на короткостроковому періоді часу.

Постановка задачі

- Шляхом моніторингу Інтернету, зокрема соціальних мереж та блогів, визначити популярність заданого бренду – тобто кількість позитивних, негативних та нейтральних його згадувань та спрогнозувати значення популярності бренду на короткостроковому періоді часу.

Актуальність

- своєчасний аналіз соціальних мереж та блогів дозволяє оперативно, максимально швидко реагувати на ситуацію, що змінилася;
- отримання загальної інформації про ситуацію на ринку, перевагах аудиторії, виявлення тенденцій і перспектив;
- об'єктивна оцінка діяльності конкурентів, їх маркетингової активності, успіхів і провалів;
- об'єктивна оцінка власної піар-активності;
- прогнозування ситуації у найближчому майбутньому враховуючи попередній досвід.

Особливості реалізації власного алгоритму

Для визначення тональності відгуків користувачів був застосований метод машинного навчання з учителем.

Процес створення системи аналізу тональності складався з таких кроків:

- Була зібрана колекція документів для навчання класифікатора (500 реальних відгуків користувачів Facebook).
- Кожен документ з навчальної колекції був представлений у вигляді вектору ознак (комбінація уніграм та біграм слів).
- Для кожного документа була вказана «правильна відповідь», тобто тип тональності (наприклад, позитивна чи негативна), за цими відповідями і був навчений класифікатор.
- В якості алгоритму класифікації був обраний наївний баєсівський класифікатор (NB).
- Використання отриманої моделі.

Алгоритм розпізнавання тональності

- Для розв’язання задачі класифікації тональності відгуків було вибрано розділення можливої тональності на 3 класи: позитивна, негативна та нейтральна.
- Спочатку був навчений бінарний класифікатор, який відокремлює нейтральні тексти від суб'єктивних; потім навчався класифікатор, який відокремлює позитивні думки від негативних.

Алгоритм розпізнавання тональності

Наївний баєсівський класифікатор — класифікатор, що використовує теорему Баєса для визначення ймовірності приналежності спостереження (елемента вибірки) до одного з класів C за умовою, що залежні змінні приймають задані значення: P (C | F1 , … , Fn).

Відповідний класифікатор - це функція classify, визначена таким чином:

$$\text{classify}(f_1, \ldots, f_n) = argmax_c\, p(c = C) \prod_{i=1}^{n} p(F_i = f_i \mid C = c)$$

- C – змінна класу;
- $F_1, \ldots, F_n$ – властивості класифікатору.

Алгоритм розв'язку задачі короткострокового прогнозування

Для короткострокового прогнозування значень часового ряду побудуємо нелінійну модель, яка дозволить відобразити розвиток експоненційної тенденції, а також рядів без тенденції:

$$Y(t+k) = A_0 + A_1 \exp(k) + A_2 \exp\left(\frac{1}{k}\right)$$

- t – поточний час;
- k – час випередження.

Цю модель було створено но основі багатьох експериментів і дана модель вперше застосовується на практиці.

Алгоритм розв'язку задачі короткострокового прогнозування

Вона будується таким чином:

1. За декількома першими точками часового ряду методом найменших квадратів знаходяться значення параметрів A_0, A_1 та A_2.

$$Y_p(t) = A_0 + A_1 \exp(t) + A_2 \exp(\frac{1}{t})$$

2. Використовуючи знайдені параметри, знаходиться прогнозне значення на наступному кроці:

$$Y_p(t+k) = A_0(t) + A_1(t) \exp(k) + A_2(t) \exp\left(\frac{1}{k}\right),\ k = 1$$

3. Знайдемо похибку прогнозування:

$$e(t+k) = Y(t+k) - Y_p(t+k)$$

4. Згідно з похибкою змінюються значення параметрів моделі:

$$A_0(t+1) = A_0(t) + A_1(t) + A_2(t) + (1-\beta)^2 e(t)$$
$$A_1(t+1) = A_1(t) + A_2(t) + (1-\beta)^2 e(t)$$
$$A_2(t+1) = A_2(t) + (1-\beta)^2 e(t)$$

β – коефіцієнт дисконтування даних, $0 < \beta < 1$.

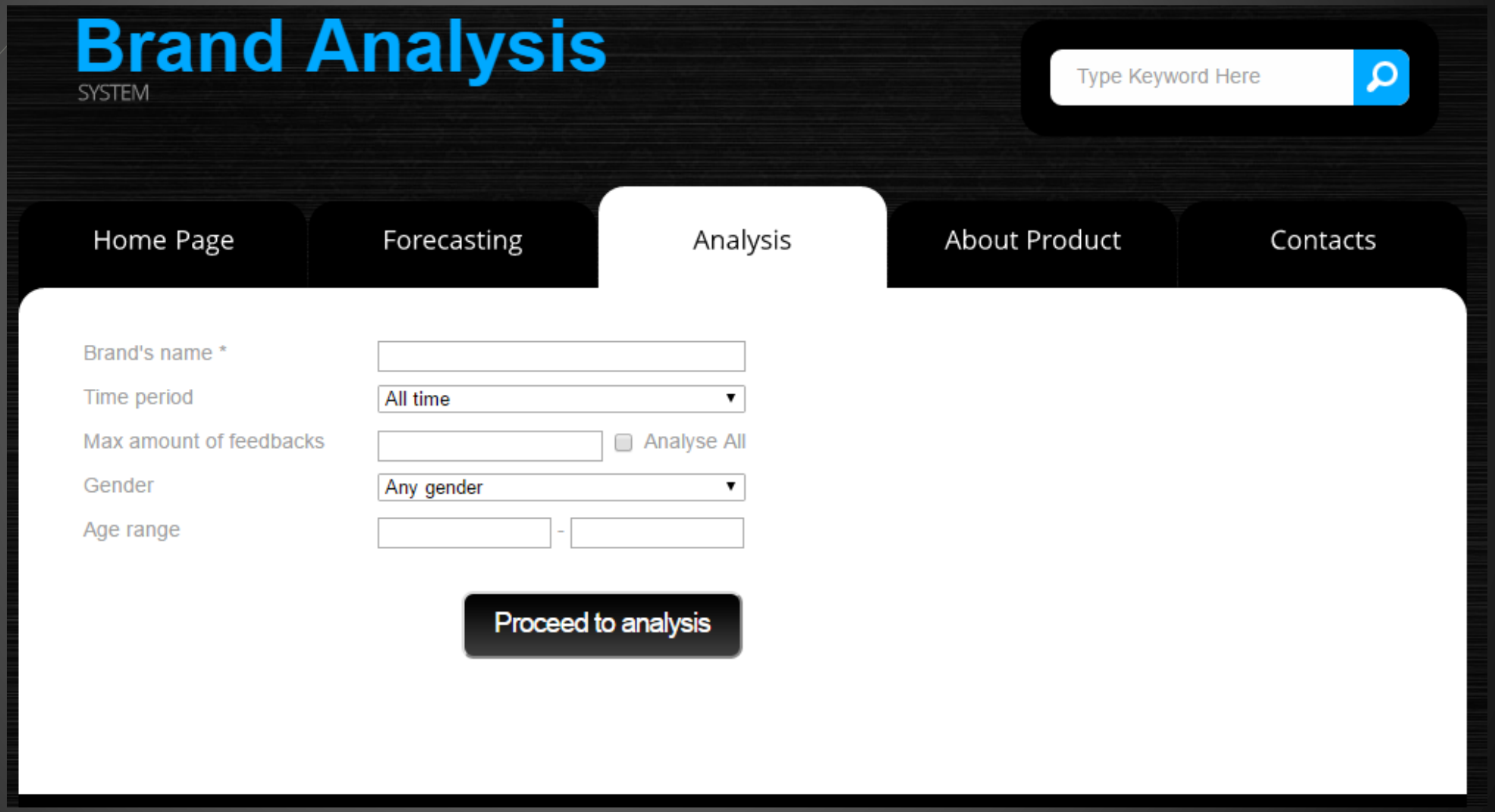
5. З моделі з виправленими параметрами знаходиться прогноз на наступний крок та відбувається перехід до п.3 якщо $t < N$. При $t \geq N$ отримане значення використовується як прогнозне, не змінюючи параметри моделі.

Brand Analysis System
Інтерфейс користувача

Програмний продукт дозволяє враховувати такі показники при аналізі брендів:

- назва бренду;
- період часу, за який користувачі згадували бренд у своїх повідомленнях (останній місяць, 2 місяці, 3, 4, 5, 6 місяців, 1 рік, 2 чи 3 роки або згадування за весь час);
- кількість відгуків, що будуть оброблюватись, та які містять згадування про бренд (будь-яке ціле число або всі відгуки);
- стать осіб, повідомлення яких будуть аналізуватися (чоловіча, жіноча або обидві);
- також можна вказати межі віку для осіб, повідомлення яких аналізується.

Інтерфейс користувача



Порівняння роботи системи моніторингу BuzzLook та власної системи

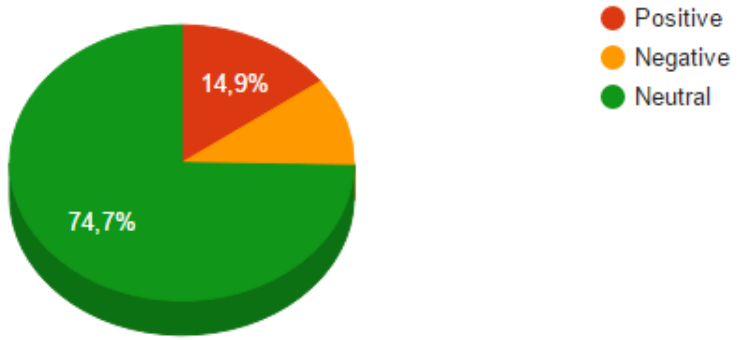
Для ілюстрування роботи двох систем та їх порівняння був обраний український бренд «Roshen». Був проведений моніторинг його згадувань за останній місяць (травень 2016 року).

Результати роботи BuzzLook

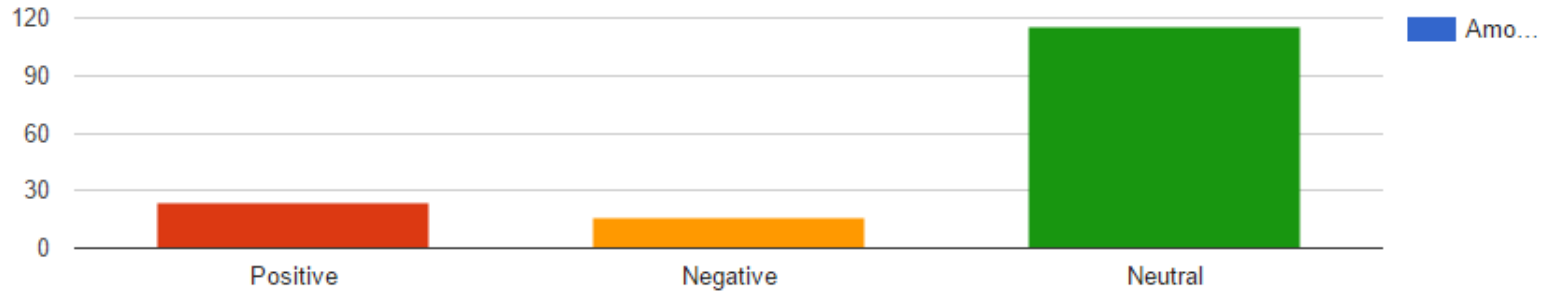
Результати роботи власної системи

All Mentions	-	154
Positive	-	23
Negative	-	16
Neutral	-	115
Undefined	-	0

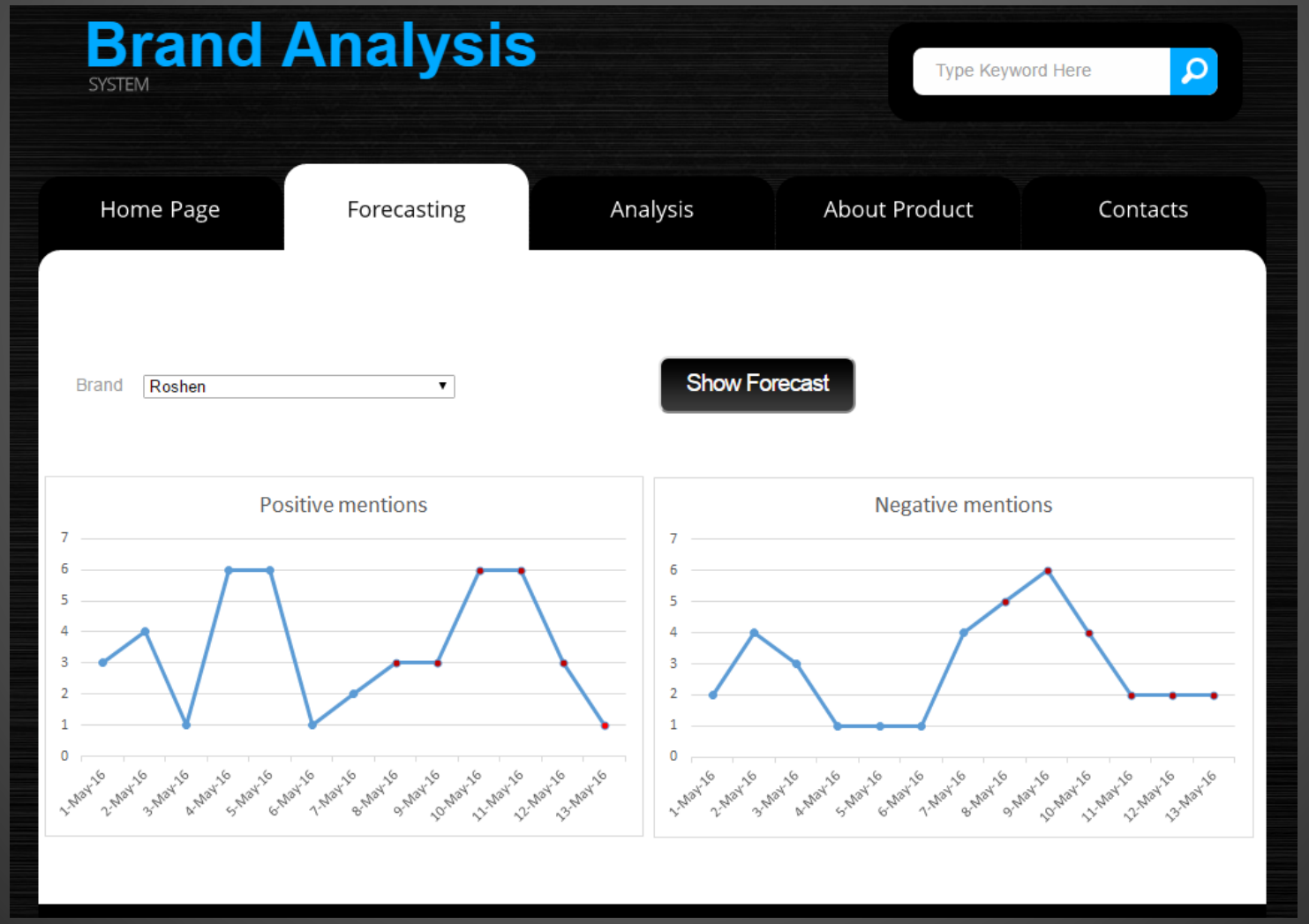
Common Statistic:



Detailed Statistics



Зпрогнозовані значення важливості бренду у власній системі



- Червоним позначені зпрогнозовані значення важливості брендів на період 6 днів.

Порівняння результатів

Показник	BuzzLook	Власна система
Кількість знайдених згадувань бренду в соціальній мережі Facebook	178	154
Кількість позитивних згадувань	0	23
Кількість негативних згадувань	0	16
Кількість нейтральних згадувань	178	115
Кількість невстановлених згадувань	0	0

Переваги та недоліки систем

BuzzLook

Переваги:

+ можливість відповідати з системи в twitter;
+ можливість проводити моніторинг більшої кількості ресурсів (twitter, facebook, vkontakte, livejournal тощо);
+ зручний інтерфейс та генерування звітів;
+ автоматичне визначення тональності відгуків.

Недоліки:

- невисока якість визначення тональності.

Власна система

Переваги:

+ можливість фільтрувати відгуки за соціальними параметрами користувачів мережі;
+ автоматичне визначення тональності відгуків;
+ можливість прогнозування популярності бренду на короткостроковому періоді;
+ доволі висока точність визначення тональностей відгуків.

Недоліки:

- проводиться моніторинг лише обмеженої кількості ресурсів: Facebook, Vkontakte, Lifehacker.ru, Businessinsider.com.

Якість рішення задачі

Тональність	Кількість відгуків, визначених системою	З них правильно визначених (думка експерта)
Позитивні відгуки	23	20
Негативні відгуки	16	15
Нейтральні відгуки	115	75

$$total\ number\ of\ opinions\ found\ by\ system = 23 + 16 + 115 = 154$$

$$correctly\ extracted\ opinions = 20 + 15 + 75 = 110$$

Визначимо точність системи за формулою:

$$P = \frac{correctly\ extracted\ opinions}{total\ number\ of\ opinions\ found\ by\ system} = \frac{110}{154} \approx 0{,}71$$

Якість рішення задачі

Згідно з дослідженням, експерти зазвичай погоджуються в оцінках тональності конкретного тексту в 79% випадків. Тобто програма, яка визначає тональність тексту з точністю 71%, робить це майже так само добре, як і людина.

Висновок

- В даній роботі було створено програмний продукт, який проводить обчислення важливості брендів у соціальних мережах та блогах та прогнозує їх важливість на короткостроковому періоді часу. Було автоматизовано встановлення тональності повідомлення, та, відповідно, аналіз важливості бренду. Для алгоритму короткострокового прогнозування була побудована нелінійна модель яка вперше використовується на практиці.
- Перспективою подальших досліджень є використання нейронної мережі, яка навчається самостійно для обчислення тональності відгуків про бренд.

Дякую за увагу!
Ваші запитання?