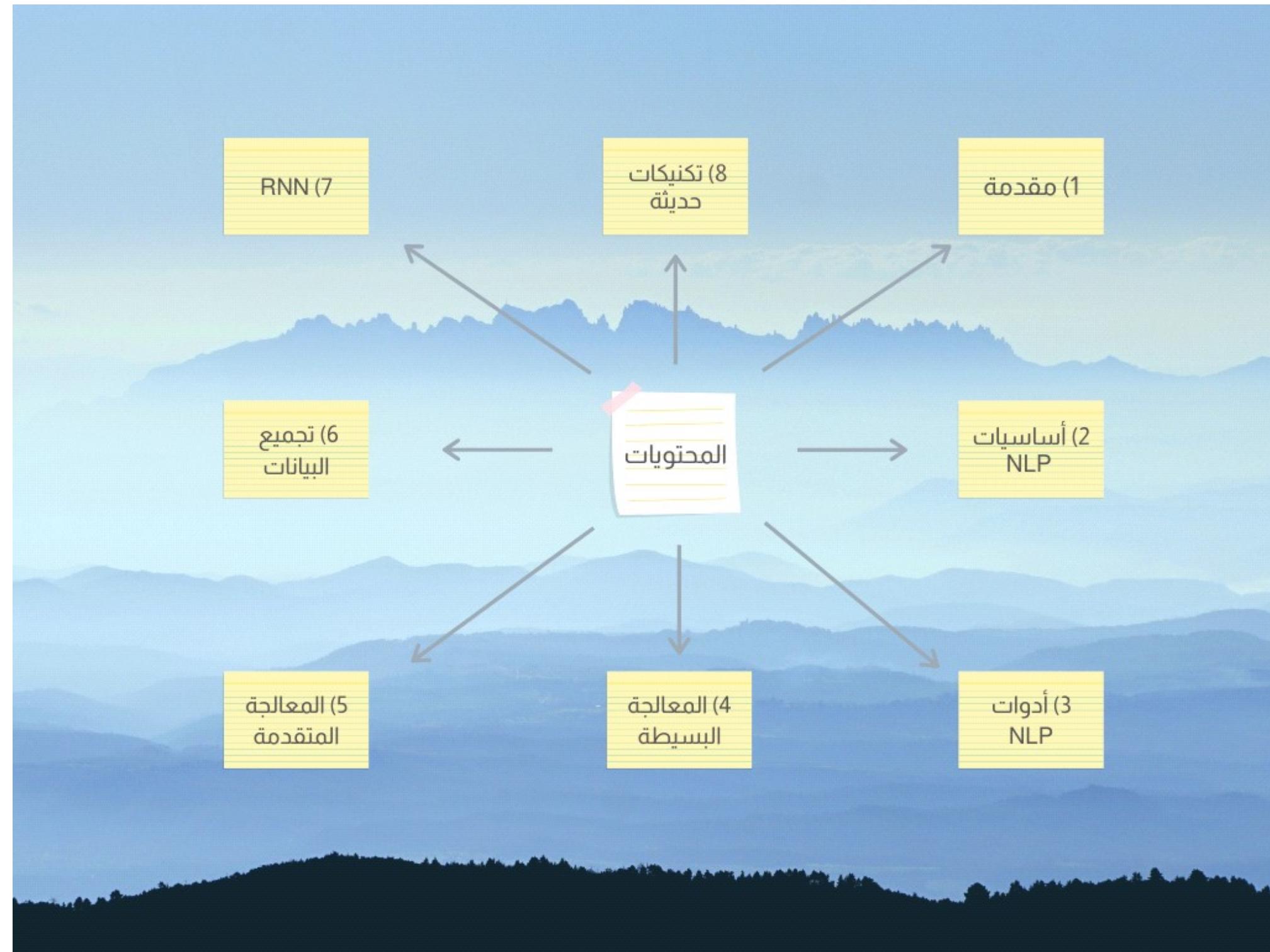


NATURAL LANGUAGE PROCESSING

المعالجة اللغوية الطبيعية



المحتويات

					التطبيقات	العقبات و التحديات	NLP تاريخ ملفات pdf	ما هو NLP الملفات النصية	المحتويات المكتبات	1) مقدمة
T.Visualization	Syntactic Struc.	Matchers	Stopwords	NER	Stem & Lemm	POS	Sent. Segm.	Tokenization	NLP أساسيات	2) أساسيات NLP
	Dist. Similarity	Text Similarity	TF-IDF	BOW	Word2Vec	T. Vectors	Word embed	Word Meaning	NLP أدوات	3) أدوات NLP
T. Generation	L. Modeling	NGrams	Lexicons	GloVe	NMF	LDA	T. Clustering	T. Classification	المعالجة البسيطة	4) المعالجة البسيطة
	Summarization & Snippets		Ans. Questions		Auto Correct	Vader	Naïve Bayes	Sent. Analysis		
Search Engine	Relative Extraction		Information Retrieval		Information Extraction		Data Scraping	Tweet Collecting	6) تجميع البيانات	
					Rec NN\TNN	GRU	LSTM	Seq to Seq	RNN (7)	
Chat Bot	Gensim	FastText	Bert	Transformer	Attention Model	T. Forcing	CNN	Word Cloud	8) تكنيكات حديثة	

القسم الخامس : المعالجة المتقدمة للنصوص

الجزء العاشر : تحليل الانطباع Sentimental Analysis

وهو من أهم تطبيقات NLP ، والذي يزداد الطلب عليه بشكل مستمر ، و له العديد من الاستخدامات في مجالات متعددة . . .

و يعرف على أنه القدرة على استخدام النصوص ، لاستخلاص معلومات هامة منها ، مثل آراء الناس و انطباعاتهم و الرأي العام و التوجه تجاه أمر ما

و قد يكون هذا مثل تحليل التقييمات الخاصة بأحد الأفلام هكذا :



- unbelievably disappointing



- Full of zany characters and richly applied satire, and some great plot twists



- this is the greatest screwball comedy ever filmed



- It was pathetic. The worst part about it was the boxing scenes.

او تقييمات مستهلكي سلعة ما ، بحيث نتمكن عبر تحليل التقييم ان نتعرف على درجات لعدد من عوامل التقييم في المنتج



HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner

\$89 online, \$100 nearby ★★★★☆ 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

Reviews

Summary - Based on 377 reviews



What people are saying

ease of use		"This was very easy to setup to four computers."
value		"Appreciate good quality at a fair price."
setup		"Overall pretty easy setup."
customer service		"I DO like honest tech support people."
size		"Pretty Paper weight."
mode		"Photos were fair on the high quality mode."
colors		"Full color prints came out with great quality."

كذلك يمكن استنتاج قيم معينة لعوامل خاصة بمنتج محدد

HP Officejet 6500A E710N Multifunction Printer

[Product summary](#) [Find best price](#) **Customer reviews** [Specifications](#) [Related items](#)



Average rating ★★★★☆ (144)

★★★★★ (55)

★★★★☆ (54)

★★★★★ (10)

★★★★★ (6)

★★★★★ (23)

★★★★★ (0)

Most mentioned

Performance

Ease of Use

Print Speed

Connectivity

More ▾

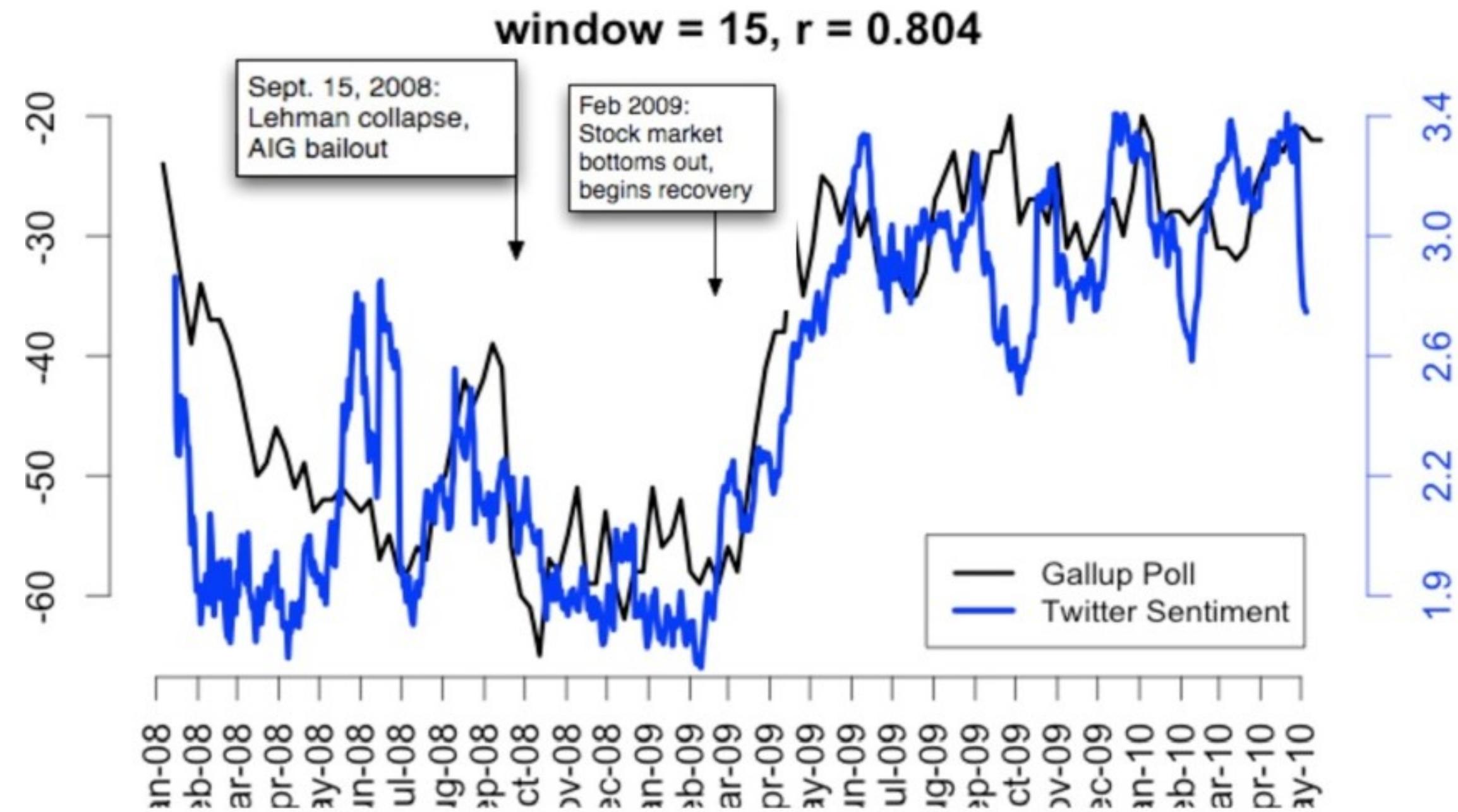
Show reviews by source

Best Buy (140)

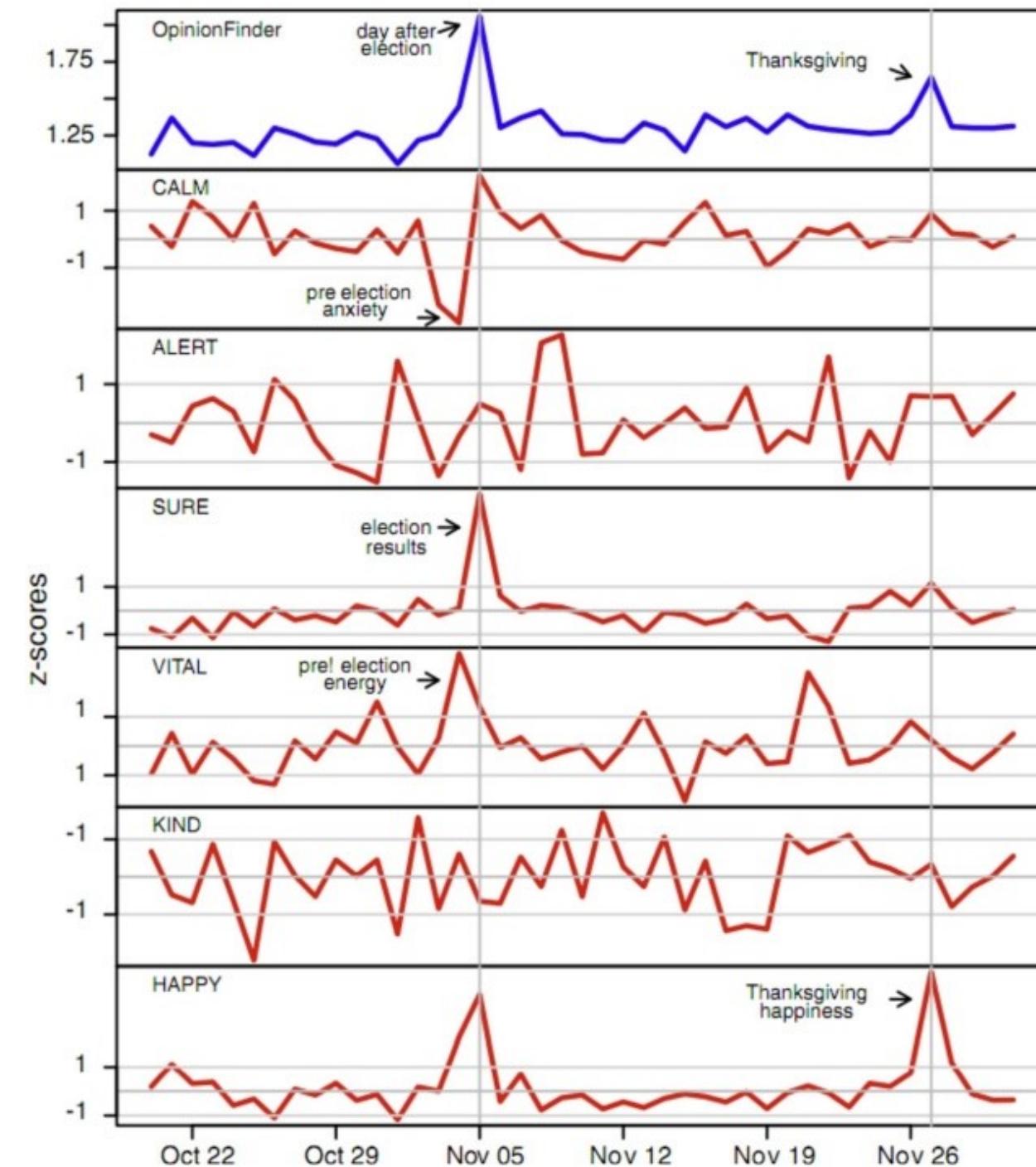
CNET (5)

Amazon.com (3)

كما ان متابعة تغريدات تويتر , يمكن لها استنتاج التوجه العام لمجموعة من المستهلكين تجاه أمر ما , فهنا نري اتجاه المستهلكين لـ "ثقة المستهلكين" في منتج معين , و نري تحليل تويتر باللون الأزرق , و بالتوازي معه تقييم شركة جالوب للمستهلكين لنفس الفترة الزمنية



و هنا مثال مشابه لتحليل عدد من المشاعر



كما ان هناك تطبيقات مختلفة لعمل تحليل للتغريدات بشكل تلقائي لفترة معينة من الزمن

Type in a word and we'll highlight the good and the bad

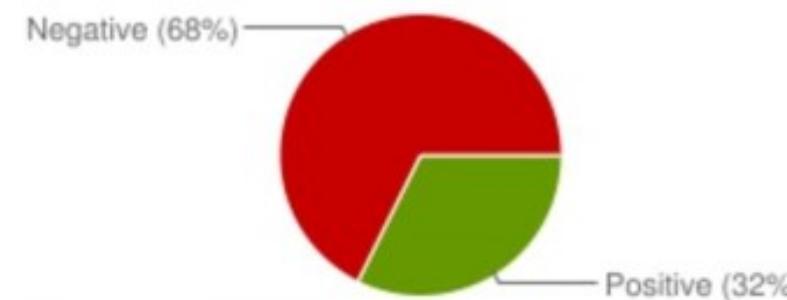
"united airlines"

Search

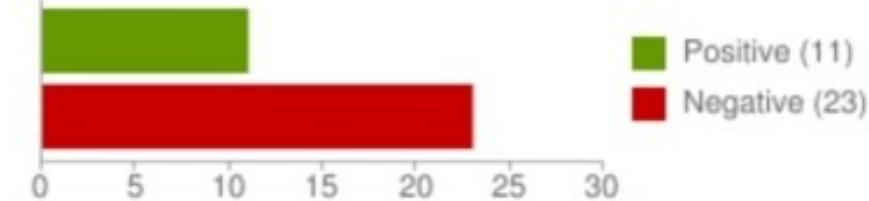
[Save this search](#)

· **Sentiment analysis for "united airlines"**

Sentiment by Percent



Sentiment by Count



jljacobson: OMG... Could @United airlines have worse customer service? W8g now 15 minut
[Posted 2 hours ago](#)

12345clumsy6789: I hate United Airlines Ceiling!!! Fukn impossible to get my conduit in this d
[Posted 2 hours ago](#)

EMLandPRGbelgiu: EML/PRG fly with Q8 united airlines and 24seven to an exotic destination
[Posted 2 hours ago](#)

CountAdam: FANTASTIC customer service from United Airlines at XNA today. Is tweet more,
[Posted 4 hours ago](#)

وهذا مثال عملی لكتمي trump & heroine

[https://www.csc2.ncsu.edu/faculty/healey/tweet viz/tweet app/](https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/)



كما ان هناك تطبيق يقوم بعمل graphs للتغريدات بشكل مستمر في مجالات مختلفة

<http://sentdex.com/financial-analysis/>

<http://sentdex.com/political-analysis/>

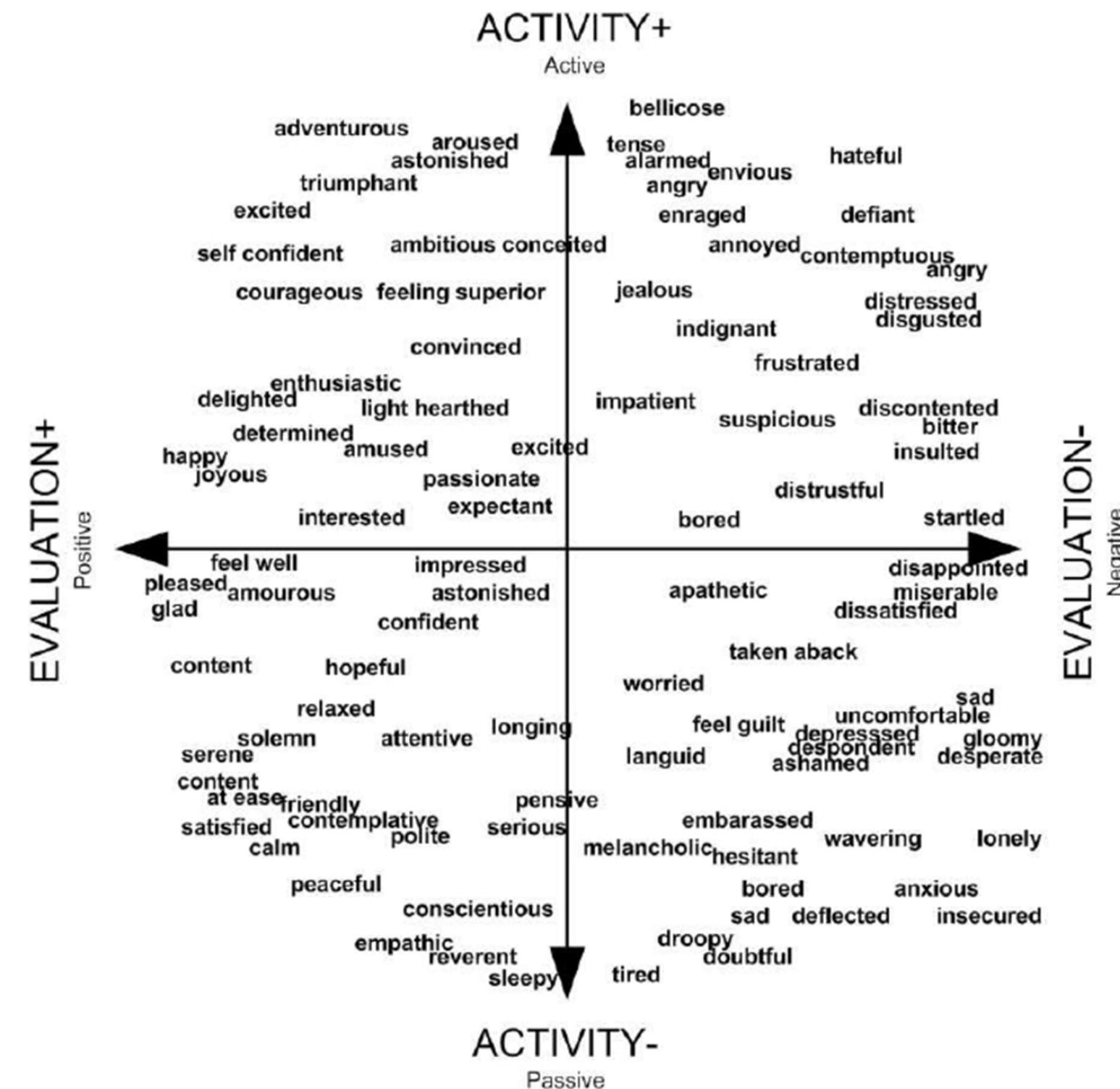
<http://sentdex.com/geographical-analysis/>

حيث تقوم هذه البرامج بجمع الكلمات و العبارات المستخدمة بشكل اوتوماتيكي من مواقع معينة (اقتصادية و سياسية و غيرها) ، وتحليلها و اظهار مدى تناولها و تكرارها و هذا

كما ان هذا الأمر له العديد من التطبيقات مثل :

- تقييم الأفلام
 - تقييم المنتجات
 - اتجاه المستهلكين
 - التطبيقات السياسية : الرأي العام
 - التوقعات , سواء في المجال السياسي او الاقتصادي او الاستهلاكي

كما ان الباحث السويسري كلاوس شيرير ، قام بعمل تصنیف لصفات الناس تبعاً لسماتهم النفسية هكذا :



و بالتالي فيمكننا ربط هذا التقسيم ، لاستخدام SA لمعرفة مود الشخص من خلال تغريداته ، و التعرف على حالته النفسية

- **Emotion:** brief organically synchronized ... evaluation of a major event
 - *angry, sad, joyful, fearful, ashamed, proud, elated*
- **Mood:** diffuse non-caused low-intensity long-duration change in subjective feeling
 - *cheerful, gloomy, irritable, listless, depressed, buoyant*
- **Interpersonal stances:** affective stance toward another person in a specific interaction
 - *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*
- **Attitudes:** enduring, affectively colored beliefs, dispositions towards objects or persons
 - *liking, loving, hating, valuing, desiring*
- **Personality traits:** stable personality dispositions and typical behavior tendencies
 - *nervous, anxious, reckless, morose, hostile, jealous*

كما انه يمكن اكتشاف أبعاد أعمق في معنى الكلام و سيكولوجية قائله هكذا :

- **Emotion:**
 - Detecting annoyed callers to dialogue system
 - Detecting confused/frustrated versus confident students
- **Mood:**
 - Finding traumatized or depressed writers
- **Interpersonal stances:**
 - Detection of flirtation or friendliness in conversations
- **Personality traits:**
 - Detection of extroverts

كما انه يمكن التعرف علي الاتجاهات السلبية و الكلام العنصري بشكل عام عبر تحليل التغريدات و الكتابات ، بل إنه يمكن التعرف علي مصدرها ، ومن المستهدف ، و نوعها

Friendly speakers use collaborative conversational style

- Laughter
- Less use of negative emotional words
- More sympathy
 - That's too bad I'm sorry to hear that
- More agreement
 - I think so too
- Less hedges
 - kind of sort of a little ...

Sentiment analysis is the detection of **attitudes**

“enduring, affectively colored beliefs, dispositions towards objects or persons”

1. **Holder (source)** of attitude
2. **Target (aspect)** of attitude
3. **Type** of attitude
 - From a set of types
 - *Like, love, hate, value, desire, etc.*
 - Or (more commonly) simple weighted **polarity**:
 - *positive, negative, neutral, together with strength*
4. **Text** containing the attitude
 - Sentence or entire document

و كان هناك ثلاث مهام :

- السهلة : تحديد هل الكلام سلبي او ايجابي
- المتوسطة : تحديد مدى سلبية او ايجابية الكلام من درجات معينة
- الصعبة : تحديد المصدر و الهدف و النوعية

--*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*

كما أنشأ علينا أن نستخرج البيانات من صفحات الويب ، لذا علينا أن نتعامل مع صفحات الـ HTML و ملفات XML ، لو تغريدات توويتر

و يمكن الابقاء على الحروف على حالتها ، أو ان نحولها جمیعا الى کابیتال او سمول ، و هذا يعتمد هل حالة الحرف ستفرق معنا في المعنی ام لا

كما ان علينا التعامل مع الارقام و التواریخ ، و حتى الحروف التعبیریة :)

- Deal with HTML and XML markup
 - Twitter mark-up (names, hash tags)
 - Capitalization (preserve for words in all caps)
 - Phone numbers, dates
 - Emoticons
 - Useful code:
 - [Christopher Potts sentiment tokenizer](#)
 - [Brendan O'Connor twitter tokenizer](#)
- Potts emoticons
- ```
[<>]?
[:::=8]
[\-o*\'']?
[(\\)]\\([dDpP/\:\:]\\){@\\}\\]
|
[(\\)]\\([dDpP/\:\:]\\){@\\}\\]
[\-o*\'']?
[:::=8]
[<>]?

optional hat/brow
eyes
optional nose
mouth
reverse orientation
mouth
optional nose
eyes
optional hat/brow
```

21

نتناول الان أحد النماذج المستخدمة في SA وهو ما يسمى : **الخط الرئيسي Base Line**

و تقوم فكرته على مراحل ثلاث :

- تحديد الكلمات
- استخراج المعلومات
- استخدام خوارزم التصنيف

و تقوم الفكرة على أن تقييم معنى الجملة ، لا يقوم على عدد تكرارات الكلمة ، ولكن فقط هل هي متواجدة أم لا ، فلو كان هناك كلمة " رائع " في الفيلم 5 مرات ، فهي مطابقة في المعنى تقريباً لتواجدها مرة واحدة ، فالملهم هل هي موجودة أم لا ، وليس كم مرة

و بالتالي فاننا نرصد عدد الكلمات مع حذف التكرار ، وهي ما يمكن ان نسميها Boolean multinomial naïve bayes

و بالتالي اذا كان لدينا الجمل السابق تحليلها , فستكون هكذا مع التناول الجديد



## Normal vs. Boolean Multinomial NB

| Normal   | Doc | Words                               | Class |
|----------|-----|-------------------------------------|-------|
| Training | 1   | Chinese Beijing Chinese             | c     |
|          | 2   | Chinese Chinese Shanghai            | c     |
|          | 3   | Chinese Macao                       | c     |
|          | 4   | Tokyo Japan Chinese                 | j     |
| Test     | 5   | Chinese Chinese Chinese Tokyo Japan | ?     |

| Boolean  | Doc | Words               | Class |
|----------|-----|---------------------|-------|
| Training | 1   | Chinese Beijing     | c     |
|          | 2   | Chinese Shanghai    | c     |
|          | 3   | Chinese Macao       | c     |
|          | 4   | Tokyo Japan Chinese | j     |
| Test     | 5   | Chinese Tokyo Japan | ?     |

علي أن لهذه الفكرة عدد من العيوب , فالتعليق الأول سلبي و الثاني ايجابي , علي الرغم من ان الكلمات هنا و هناك لا تعبر اطلاقا عن هذا

## Subtlety:

- Perfume review in *Perfumes: the Guide*:
  - “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”
- Dorothy Parker on Katherine Hepburn
  - “She runs the gamut of emotions from A to B”

كما ان التعليقات التي تعبّر عن احباط صاحبها من الفيلم ستكون مشكلة في تقييمها ، ففي التعليقات كلمات ايجابية زرقاء و سلبية حمراء ، ومن الصعب على الموديل فهم المعنى الحقيقي عبر البحث عن الكلمات

- “This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up.”
  - Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is not so good either, I was surprised.

لذا فإن هناك خوارزميات أخرى يمكن لها أداء قد يكون أفضل مثل MaxEnt , SVM , Lexicons



كما أن خوارزم SA لا يكتفي فقط بإعطاء درجة الجملة ، ولكن يستخلص منها معلومات أكثر . .

فلو كان لدينا هذه الجملة :

The food was great , but the service was awful

، فهي تعني شيء إيجابي و شيء سلبي

لكن الأهم أن علينا أن ندرك أن الجانب الإيجابي في نقطة معينة ، والجانب السلبي في نقطة أخرى ، فهذه الجملة تختلف عن جملة :

The food was awful , but the service was great

فعلي الرغم أن الجملتين فيهما نقطة إيجابية و أخرى سلبية ، لكن من المهم أن ندرك أن كلاً منهما تمدح في شيء و تذم في شيء آخر . .

و من المهم أن يتم استخراج إيجابيات و سلبيات كل تعليق ، وليس فقط استخراج الدرجة الكلية

و يتم هذا عبر البحث عن كلمات دلالية معينة في التعليقات ، وهذه الكلمات تكون خاصة بنوع التعليقات ، هل هي أفلام أم مطعم أو فندق ، حيث تكون الكلمات الدلالية هي التي تحدد نوع الخدمة داخل التعليق

- Frequent phrases + rules
  - Find all highly frequent phrases across reviews (“fish tacos”)
  - Filter by rules like “occurs right after sentiment word”
    - “...great fish tacos” means fish tacos a likely aspect

|                   |                                              |
|-------------------|----------------------------------------------|
| Casino            | casino, buffet, pool, resort, beds           |
| Children’s Barber | haircut, job, experience, kids               |
| Greek Restaurant  | food, wine, service, appetizer, lamb         |
| Department Store  | selection, department, sales, shop, clothing |

و يمكن أن تكون الكلمة غير موجودة في التعليق (مثلا تعليق عن مطعم و لم يكون فيه كلمة food), لذا علينا ان نبحث عن جميع الكلمات المتعلقة بهذا الأمر ، حتى لو كانت كلمة قريبة منها

## Rooms (3/5 stars, 41 comments)

- (+) The room was clean and everything worked fine – even the water pressure ...
- (+) We went because of the free room and was pleasantly pleased ...
- (-) ...the worst hotel I had ever stayed at ...

## Service (3/5 stars, 31 comments)

- (+) Upon checking out another couple was checking early due to a problem ...
- (+) Every single hotel staff member treated us great and answered every ...
- (-) The food is cold and the service gives new meaning to SLOW.

## Dining (3/5 stars, 18 comments)

- (+) our favorite place to stay in biloxi.the food is great also the service ...
- (+) Offer of free buffet for joining the Play

و علينا أن نتأكد أن هناك توازن في البيانات (عدد متقارب بين التعليقات السلبية والإيجابية) و إذا لم يكن هناك توازن، فيمكن حل الأمر بـ :

- الاعتماد على F Score وليس ال accuracy لأن F Score تعتمد على مجموعة  $TP$ ,  $TN$  بناء على الجميع
- استخدام فكرة random undersampling اي لو كان لدينا مليون تعليق سلبي و 10 الاف تعليق ايجابي ، فيتم اختيار من المليون تعليق سلبي عدد 10 الاف بشكل عشوائي
- استخدام cost function اي خوارزميات تقوم بزيادة ال cost sensitive learning كلما كانت الفئة المستخدمة اقل ، وهذا سيجعل الخوارزم يقوم بعمل الموازنة تلقائيا

و اذا كان لدينا التقييم ليس ( سلبي ايجابي) و لكن من درجات ( 5 او 7 درجات ) فمن الممكن ان يتم تحويلها الى تصنيف ثنائى ( الدرجة اقل من 2.5 تكون 0 و اكبر تكون 1 )

او ان يتم استخدام regression او multi classification

\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*

و من مبادئ الـ sentimental analysis : منها ما يسمى positive and negative counts :

فلو كان لدينا مجموعة من الجمل مثل :

I am happy because I am learning NLP

I am happy

I am sad I am not learning NLP

I am sad

فلو كانت الجملتين الاولى و الثانية بمعنى ايجابي و الثالثة و الرابعة بمعنى سلبي فاولا يتم حصر جميع الكلمات وهي :

I , Am , happy , because , learning , NLP , Sad , Not

ثم يتم فحص جميع الجمل الايجابية ، و عمل عد لاي كلمات متواجدة فيها ، ثم عمل العكس في الجمل السلبية ، للحصول على

PosFreq & NegFreq

| Vocabulary | PosFreq (1) | NegFreq (0) |
|------------|-------------|-------------|
| I          | 3           | 3           |
| am         | 3           | 3           |
| happy      | 2           | 0           |
| because    | 1           | 0           |
| learning   | 1           | 1           |
| NLP        | 1           | 1           |
| sad        | 0           | 2           |
| not        | 0           | 1           |

ثم نقوم بالحصول على عدد التكرارات في كل تويبة ( او جملة) اعتمادا علي الجدول السابق

*freqs*: dictionary mapping from (word, class) to frequency

$$X_m = [1, \sum_w freqs(w, 1), \sum_w freqs(w, 0)]$$

↓              ↓              ↓              ↓  
 Features of    Bias          Sum Pos.      Sum Neg.  
 tweet m                     Frequencies      Frequencies

و بالتالي مع هذه الجملة , سنري ان الكلمات المستخدمة في الجملة

I am sad I am not learning NLP

## Feature extraction

| Vocabulary | PosFreq (1) |
|------------|-------------|
| I          | <u>3</u>    |
| am         | <u>3</u>    |
| happy      | 2           |
| because    | 1           |
| learning   | <u>1</u>    |
| NLP        | <u>1</u>    |
| sad        | <u>0</u>    |
| not        | <u>0</u>    |

I am sad, I am not learning NLP

$$X_m = [1, \sum_w freqs(w, 1), \sum_w freqs(w, 0)]$$

8

## Feature extraction

| Vocabulary | NegFreq (0) |
|------------|-------------|
| I          | <u>3</u>    |
| am         | <u>3</u>    |
| happy      | 0           |
| because    | 0           |
| learning   | <u>1</u>    |
| NLP        | <u>1</u>    |
| sad        | <u>2</u>    |
| not        | <u>1</u>    |

I am sad, I am not learning NLP

$$X_m = [1, \sum_w freqs(w, 1), \sum_w freqs(w, 0)]$$

11

لذا فالعدد النهائي هو :

I am sad, I am not learning NLP

$$X_m = [1, \sum_w freqs(w, 1), \sum_w freqs(w, 0)]$$

↓

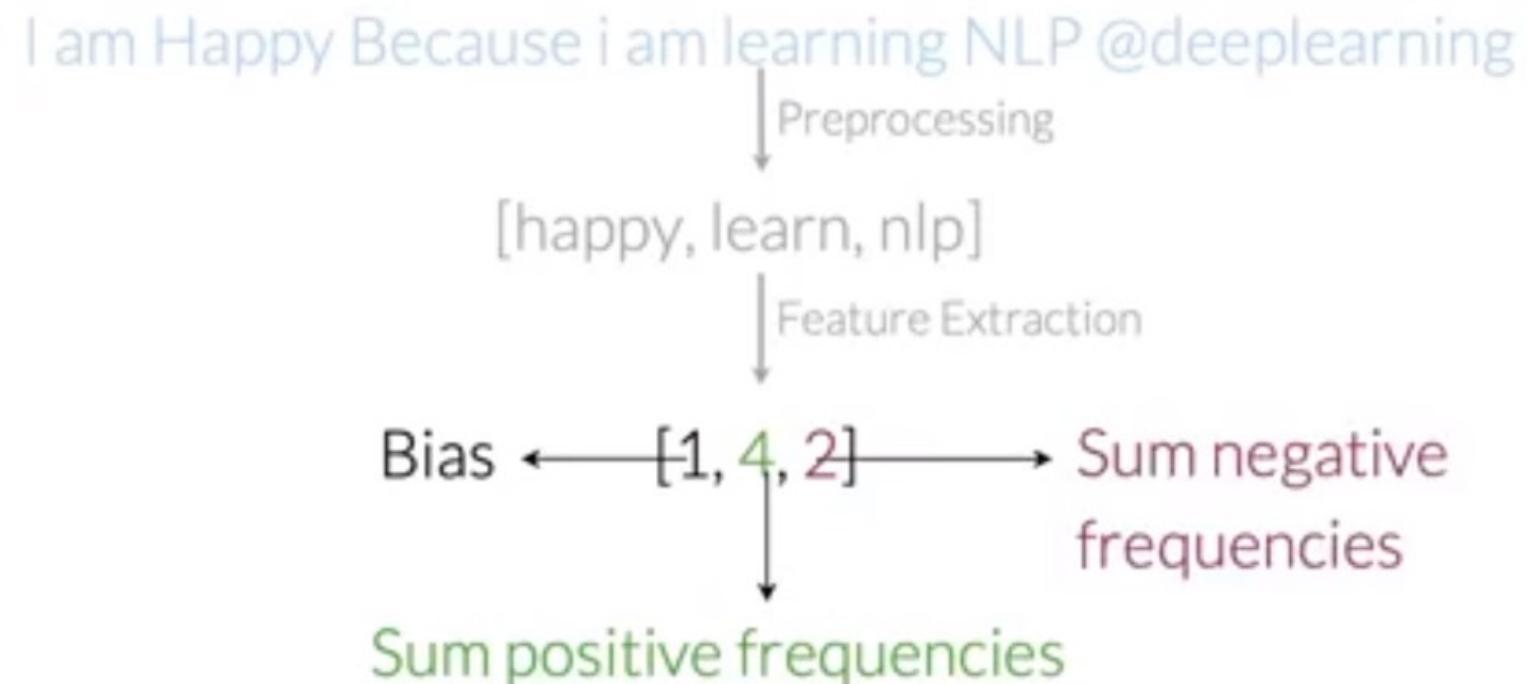
$$X_m = [1, 8, 11]$$

و كل هذا يأتي بعد مرحلة التخلص من جميع الاضافات ، فالstopwords & stemming ، هي التخلص من stopwords معرفة و يتم التخلص منها جمیعا او بعضها

و لا تنس أن الـ stemming مقصود بها ارجاع الكلمة لأصلها , فجميع كلمات plays , playing , played , player ترجع لكلمة play

و بالتالي العملية ستكون كالتالي :

ولا تنظيف البيانات و اعدادها , و عمل اعداد الايجابي و السلبي



ثم تطبيق هذا العد على جميع الجمل الموجودة في البيانات

|                                                          |                        |                                       |
|----------------------------------------------------------|------------------------|---------------------------------------|
| I am Happy Because i am<br>learning NLP<br>@deeplearning | [happy, learn, nlp]    | [[1, 40, 20],<br>[1, 20, 50],<br>...] |
| I am sad not learning NLP                                | [sad, not, learn, nlp] | [[1, 5, 35]]                          |
| ...                                                      | [sad]                  | ...                                   |
| I am sad :(                                              |                        |                                       |

ثم عمل مصفوفة عامة ، تكون 3 اعمدة ، وبها عدد صفوف يساوي عدد الجمل المطلوبة ، بحيث يكون العمود الاول ارقام 1 ( باب ) و العمود الثاني و الثالث ارقام الموجب و السالب

$$\begin{bmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ 1 & X_1^{(2)} & X_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & X_1^{(m)} & X_2^{(m)} \end{bmatrix} \longleftrightarrow \begin{bmatrix} [1, 40, 20] \\ [1, 20, 50], \\ \dots \\ [1, 5, 35] \end{bmatrix}$$

و عبر التعامل مع هذه المصفوفة ، يمكن استخلاص المعلومات ، او عمل تصنیف للنصوص

و بالنسبة للغة العربية، فيمكن عمل التطبيق بشكل متشابه اذا توافرت الداتا المعونة المناسبة