Opinion mining

3- a review of nlp tecch for om systems

1-om and information fusion :survey

2-a survey on om and sa:task..

ont

.

# سطوح تحلیل

در زمان توسعه یOpinion mining در ابتدا در سطح داکیومنت انجام شد و هدف آن ‍ پیدا کردن گرایش کلی داکیومنت بود.سپس این فرایند بر روی سطح جمله انجام شد و در نهایت در سطح موجودیت و aspectها. تحلیل هایی که ریز دانه ترند می توانند تجمیع شوند تا سطوح بالاتر را بسازند. برای مثال می توان از روی عقیده کاوی aspect-based مقدار متوسط احساسات را در یک جمله برای به دست آوردن نتایج در سطح جمله محاسبه کرد.

سطح داکیومنت: عقیده کاوی در این سطح تحلیل تلاش می کند تا یک داکیومنت را به کلاس های مثبت یا منفی دسته بندی کند. کاربرد این سطح معمولا review analysis است.(بررسی فیلم، نظرات محصول، توییت و بلاگ ها)

A review of natural language processing techniques for opinion

mining systems

وظیفه ی این مرحله استخراج گرایش احساس کلی متن است.این مرحله فرض می کند که تنها یک target(هدف) عقیده در یک داکیومنت بررسی شده.

سطح جمله: این سطح معادل مورد قبلی است چون می توان یک جمله را یک داکیومنت کوتاه در نظر گرفت.

مشابه سطح داکیومت وظیفه سطح جمله نیز مسئله ی کلاس بندی است اما بر روی هرجمله ی داکیومت تمرکز می کند و مشخص می کند که یک جمله گرایش مثبت منفی یا خنثی ابراز می کند.

کلاس بندی Subjectivity نیز یکی دیگر از وظایف سطح جمله است که جملات Subjective را از بخش های objective داکیومنت جدا می کند. ~~این وظایف به استخراج جزئیات مانند تارگت عقیده یا فردی که آن نظر را ابراز کرده توجهی ندارند~~

~~برای مثال در review صفحه نمایش این موبایل خوب است سه مورد اول به صورت زیر تعیین می شوند~~

~~موجودیت موبایل استaspect صفحه نمایش و عقیده ی مثبت بیان شده است. وظایف سطح ویژگی با این مسائل فراتر از تکنیک های کلاس بندی سروکار دارد.~~

Opinion Mining and Information Fusion: A survey

سطح ویژگی: این سطح ریزدانه ترین سطح است که در آن عقیده کاوی انجام می شود. در اینجا ،وظیفه، تنها پیدا کردن گرایش عقیده نیست بلکه یافتن target (موجودیت ، جنبه یا هردو)نیز هست. فرایند انجام عقیده کاوی در سطح ویژگی ابتدا تشخیص ویژگی های محصول و سپس تشخیص عقیده هایی که به این ویژگی ها اشاره کرده اند می باشد. سپس تعیین گرایش آن ها ~~و در نهایت رتبه بندی آن ها بر اساس قدرتشان.~~

A review of natural language processing techniques for opinion

mining systems

برای مثال در مورد نظرات یک محصول، موجودیت به طور صریح مشخص است(برابر نام محصول) و چندین ویژگی این موجودیت در عقیده ها ذکر شده اند. هدف در این مرحله پیدا کردن target خاص(جنبه یا موجودیت) و گرایش احساس متناظر با آن است.

از سطح داکیومنت تا سطح ویژگی پیچیدگی مسائل افزایش میابد. نبود داده های برچسب دار در سطح های دقیق تر آن را سخت تر نیز می کند. معمولا روش های supervised بهتر از روش های unsupervised عمل می کنند اما کمبود منابع تحقیقات را به سمت توسعه ی روش های semi supervised یا unsupervised مایل کرده است.

# Data acquisition

Opinion Mining and Information Fusion: A survey

گام اول برای فرایند عقیده کاوی جمع آوری داده ها است و شامل به دست آوردن مجموعه داده هایی است که می خواهیم برای یافتن عقیده ها کاوش کنیم. دو روش برای انجام این کار وجود دارد.

1- از طریق Application Programming Interface سایت ، مانند تویتر

2- استفاده از web crawler ها برای به دست آوردن داده ها از سایت های مورد نظر

هردو این روش ها فواید و مشکلاتی دارند و trade-off های بین استفاده از هرکدام از آن ها وجود دارد

با استفاده از روش مبتنی بر API پیاده سازی آُسان است و احتمال اینکه ساختار داده ی جمع آوری شده تغییر کند کم است.اما این محدودیت هایی را بر اساس فراهم کننده ی API ایجاد می کند . مثلا سرعت خواندن داده ها توسط کلاینت در تویتر محدود است. این روش البته نیاز به وجود این API ها در آن وب سایت دارد و تمامی وب سایت ها آن را فراهم نمی کنند و حتی در صورت وجود آن ممکن است همه ی کاربرد های مورد نیاز موجود نباشد.

پیاده سازی روش های crawler-based سخت تر است چون داده های به دست آمده نویزی تر بوده و ساختار آن ممکن است تغییر کند اما این روش ها مزیت نامحدود بودن را دارند.

# preprocess

123

معمولا داده ی خام جمع آوری شده از منابع مختلف پیش از انجام تحلیل کامل نیاز به پیش پردازش دارد. برخی گام های پیش پردازش عبارت است از:

tokenizationنشانه گذاری

برای تبدیل کردن جمله به کلمات و عبارات و نماد ها یا دیگر توکن های معنادار استفاده می شود و در زبان انگلیسی با جداکردن لغات به وسیله ی space ها به آسانی انجام می شود.

stop word removal,

فرایند حذف کلماتی که برای ساختاردهی زبان استفاده می شوند اما در محتوای آن نقشی ندارند. مانند a, are, the, was و will

Stemmingریشه یابی

پروسه ی تبدیل کردن کلمه به حالت ریشه ای آن با حذف پسوند و پیشوندهای کلمه و تبدیل به شکل اصلی یا ریشه ی آن

برای مثال person, person’s, personify ,personification تبدیل بهperson می شوند.

parts of speech (POS) tagging, برچسب گذاری نقش کلمه

این مرحله برای تشخیص نقش کلمات در متن انجام می شود و در عقیده کاوی بسیار کمک کننده اند چون کلمات عقیده ها معمولا صفت بوده و و تارگت عقیده ها(موجودیت ها و جنبه ها) معمولا اسم یا ترکیبی از اسامی هستند و این برچسبها می توانند برای استفاده به عنوان features برای فرایند machine learning استفاده شوندمانند(؟)adjective, noun, verb,

adverb and preposition

همه ی فرایندهای گفته شده همیشه ضروری نیستند و باید بر اساس کاربرد عقیده کاوی برای مسئله انتخاب شند(مثل؟)

unsupervised approach might need some of the stopwords’

parts of speech to build the dependency rules later used in the

Opinion Mining core process [43] therefore omitting the stopword

removal process.

راهبرد های مختلف:

Opinion Mining and Information Fusion: A survey

دو روش اساسی برای انجام فرایند اصلی عقیده کاوی وجود دارد .

روش unsupervised lexicon-based :

نام دیگر آن semantic-based است ودر این روش، تلاش می کند تا بر اساس قوانین وابسته‌ به‌ زبان‌ شناسي‌ گرایش متن را استخراج می کند.

Ont

این روش بر اساس منابع دانش خارجی عمل می‌کند و بر خلاف روش supervised شامل مرحله ی آموزش نیست.

گام های این روش به طور کلی به صورت زیر است:

1. استخراج کلمات یا عباراتی که گرایش احساسی متن را بیان می کنند.
2. تعیین گرایش این عبارات استخراج شده.
3. تعیین گرایش متن با اجتماع گرایش تک تک کلمات.

1

روش supervised learning-based یا machine-learning-based که در آن اطلاعات زیربنایی را از داده های آموزشی یاد می گیرد که به آن امکان کلاس بندی داده های بدون برچسب جدید را می دهد.

این روش معمولا شامل مهندسی feature ها برای نمایش object ی که باید کلاس آن تعیین بشود و سپس استفاده از این نمایش به عنوان ورودی الگوریتم می باشد. برخی از feature هایی که در عقیده کاوی به کار می رود عبارت است از:

تعداد تکرار کلمات، نقش کلمات، کلمات یا عبارات احساسی و...

همچنین تعداد فزاینده ای از مطالعات بر روی ترکیب موفق این دو روش وجود داشته است.

الگوریتم ها

(سطح-نوع؟از کجا اومدش sentence level sentiment classification)

S.-M. Kim , E. Hovy , Determining the sentiment of opinions, in: Proceed- ings of the 20th

International Conference on Computational Linguistics, 2004, pp. 1367–1373 .

در این روش با داشتن موضوع( به عنوان ورودی) و مجموعه ای از متن ها راجع به موضوع، احساسات اظهار شده در مورد موضوع در هر متن را به دست می آورد و افرادی که هر احساس را بیان کرده اند را تشخیص می دهد. احساسات می تواند مثبت یا منفی یا خنثی باشند. در این روش از از کلمات به سمت جمله حرکت می کنیم. هر کلمه را به عنوان واحد نشان دهنده ی احساسات در نظر گرفته و در ابتدا کلاس هر کلمه را بر اساس احساسات آن دسته بندی می کنیم.

هر جمله ممکن است حتی بیانگر نظرات چند نفر باشد و هنگام ترکیب احساسات در سطح کلمات، به همین دلیل ما ابتدا برای نظرهنده ه های مختلف بخش مرتبط با او را در جمله مشخص می کنیم و سپس از مدل های مختلف برای ترکیب احساسات کلمات استفاده می کنیم.

با داشتن موضوع و مجموعه ای متن ها سیستم در چهارگام این عملیات را انجام می دهد. اول جمله هایی که شامل عبارت موضوع و کاندیدا های افرادی که نظر را داده اند را انتخاب می کند. سپس مرز ناحیه ی مبتنی بر نظر دهنده را مشخص می کند. پس از آن طبقه بند احساسات جملات گرایش تک تک کلمات دارای احساسات را محاسبه می کند و در آخر سیستم آن ها را برای تولید احساسات نظر دهنده برای کل جمله را ترکیب می کند.

Miller, G.A., R. Beckwith, C. Fellbaum, D.Gross, and K. Miller. 1993. Introduction to WordNet: An On-Line Lexical Database. http://www.cosgi.princeton.edu/~wn.

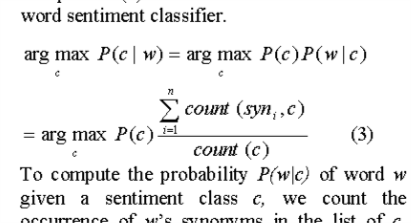
WordNet یک مرجع آنلاین لغات است که در آن اسم ها و فعل ها و صفت های انگلیسی در گروه های مترادف قرار دارند.

word sentiment classification

در روش اصلی برای به دست آوردن احساسات کلمات ، مجموعه ی کوچکی از کلمات را به صورت دستی جمع آوری می کنیم و بر اساس گرایش به دو دسته تقسیم می کنیم.(مثبت ومنفی) سپس این مجموعه را با اضافه کردن کلمات مترادف و متضاد آن ها از WordNet گسترش می دهیم با فرض اینکه اکثر مترادف های یک کلمه مثبت هم مثبت بوده و متضاد های آن منفی هستند .

برخی کلمات مانند great, strong , take چندین با در دسته های مثبت و منفی دید شدند که نشان می دهد نیاز به معیار قدرت احساسات گرایش هر کلمه داریم. این روش به ما امکان این را می دهد که کلماتی که از لحاظ احساسی مبهم هستند را دور انداخته و تنها کلماتی که قدرت آن ها از مقدار آستانه بیشتر است را نگه داریم.

با این روش می توانیم قدرت احساسی کلماتی که تا کنون مشاهده نشده اند را نیز تعیین کنیم و به این صورت عمل می کنیم که با استفاده از WordNet مجموعه کلمات مترادف آن کلمه جدید را تعیین می کنیم و به صورت زیر عمل می کنیم:



که در آن C دسته بندی احساسات است(مثبت و منفی) و w کلمه ی جدید است. synn کلمات مترادف w استخراج شده از WordNet اند.

در این روش برای محاسبه ی P(w|c) کلمه ی w با داشتن کلاس c ، تعداد رخ دادن مترادف های w را در لیست c می شماریم ( ایده اصلی این است که هرچه بیشتر مترادف ها در این کلاس باشند، با احتمال بیشتری این کلمه عضو این کلاس است). برای هرکلمه قدرت مثبت و منفی را محاسبه می کنیم و قدر مطلق آن ها را مقایسه می کنیم.

مثال:

abysmal : NEGATIVE

[+ : 0.3811][- : 0.6188]

adequate : POSITIVE

[+ : 0.9999][- : 0.0484e-11]

afraid : NEGATIVE

[+ : 0.0212e-04][- : 0.9999]

ailing : NEGATIVE

[+ : 0.0467e-8][- : 0.9999]

amusing : POSITIVE

[+ : 0.9999][- : 0.0593e-07]

answerable : POSITIVE

[+ : 0.8655][- : 0.1344]

apprehensible: POSITIVE

[+ : 0.9999][- : 0.0227e-07]

averse : NEGATIVE

[+ : 0.0454e-05][- : 0.9999]

blame : NEGATIVE

[+ : 0.2530][- : 0.7469]

Sentence Sentiment Classifier

ما در جستوجوی احساساتی هستیم که نظر دهنده ابراز کرده است تحلیل ها نشان داده اند که این احساسات در جمله نزدیک به نظردهنده به طور قابل اطمینان تری به دست می آیند. به همین دلیل در این مراحل ، گام هایی برای یافتن موضوع با جست و جوی خود کلمه در متن و هر نظردهنده ی ممکن اضافه شده است. سپس نزدیک هر نظر هنده منطقه ای که احساسات آن باید در نظر گرفته شود را مشخص کنیم.

تشخیص نظردهنده

Named entity یک مفهوم در دنیای واقعی است مانند افزاد مکان ها سازمان ها محصولات و..که می تواند با یک نام خاص نشان داده شود

Named-entity recognition در یک متن بدون ساختارمحل Named entity ها را یافته و آن ها را در کلاس های از قبل مشخص شده، کلاس بندی می کند .

در اینجا از BBN’s named entity tagger IdentiFinder برای تشخیص نظردهندگان یک عقیده استفاده شده است و فقط فرد یا سازمان را به عنوان نظردهنده های ممکن در نظر گرفته است. در جمله هایی با بیشتر از یک نظر دهنده، موردی که به عبارت موضوع نزدیکتر بود را انتخاب می کند.

خود موضوع رو مشخص می کند؟

Sentiment Region

به صورت های مختلفی می توان ناحیه ای که احساسات مورد نظر در آن بیان شده را مشخص کرد

پنجره 1: کل جمله

پنجره 2: لغات بین نظردهنده وعبارت موضوع

پنجره 3: همان پنجره2 ± دوکلمه

پنجره 4: همان پنجره 2 تا پایان جمله

مدل های کلاس بندی:

برای انتساب یک کلاس احساسی به جمله ی داده شده می توان از راه های زیر استفاده کرد:

مدل 1:

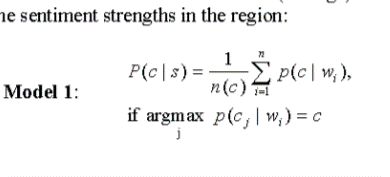
Π (signs in region)

ایده ی این نظر مانند این است که منفی در منفی برابر مثبت است. در این مدل کلماتی مانند not و neverمی تواند گرایش احساسی را وارونه کند. این روش به این دو جمله یک احساس را نسبت می دهد:

"The California Supreme Court ***agreed*** that the state’s new term-limit law was ***constitutional***”

“the California Supreme Court ***disagreed*** that the state’s new term-limit law was ***unconstitutional***”

مدل دوم میانگین قدرت احساسات در جمله است



در این مدل n( c) تعداد کلمات در ناحیه است که دسته بندی احساسی آن برابر c است.در صورتی که جمله ای دارای کلمات مثبت بیشتر و قوی تری باشد ، جمله مثبت خواهد بود.

مثال:

For that reason and others, the Constitutional

Convention unanimously rejected term limits

and the ***First Congress*** soundly defeated two

subsequent ***term-limit*** proposals.

TOPIC : term limit

HOLDER : First Congress

OPINION REGION: soundly/RB defeated/VBD

two/CD subsequent/JJ term-limit/JJ

proposals./NN

SENTIMENT\_POLARITY: negative

W. Medhat et al., Sentiment analysis algorithms and applications: a survey,

Ain Shams Eng. J. (2014), http://dx.doi.org/10.1016/j.asej.2014.04.011.

Naı¨ve Bayes Classifier

Naı¨ve Bayes ساده ترین و معمول ترین classifier. است. این مدل احتمال پسین هر کلاس(مثبت یا منفی بودن عقیده) را بر اساس پراکندگی کلمات در داکیومنت به دست می آورد و براساس روش استخراج ویژگی bag of words کار می کند که محل کلمات در داکیومنت را نادیده می گیرد.

~~این روش از نظریه بیز برای پیش بینی احتمال اینکه برچسب مخصوص برای ویژگی های داده شده کدام است استفاده می کند.~~

Opinion Mining using Naïve Bayes

کلاس داکیومت برابر برچسب مربوط به بیشترین احتمال زیر است.

B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, Association for Computational Linguistics, 2002, pp. 79–86.

P(label|document)=

با فرض مستقل بودن رخ دادن کلمات(ویژگی ها) از یکدیگر، عبارت بالا به صورت زیر تبدیل می شود:

P(label|document)=

P(document) تاثیری در انتخاب label ندارد پس می توان آن را حذف کرد

مرحله ی آموزش محاسبه ی P(label) :

P(label)=

و P(fi|label) با استفاده از add-one smoothing به شکل زیر خواهد بود:

P(fi|label)=

با این روش به کلماتی که تا کنون مشاهده نشده اند احتمالی اختصاص داده می شود و جمع احتمال همه ی کلمات در کلمات یک برچسب همچنان یک باقی خواهد ماند.(منبع؟)

~~Opinion Mining using Naïve Bayes~~

~~P(feature|label) احتمال اولیه این است که مجموعه ویژگی های داده شده با برچسب label کلاس بندی شوند.~~

~~P(features) احتمال اولیه آن است که مجموعه ویژگی های داده شده رخ دهد.~~

مثال

برای آموزش سیستم داکیومنت هایی مانند نظرات راجع به محصولات (مانند ساعت)که دارای امتیاز نیز هستند انتخاب می شود و امتیاز های بالا(4و5) به برچسب مثبت و امتیاز های پایین(1و2) به برچسب منفی تبدیل شد Review ها با امتیاز 3 به عنوان نمونه های خنثی حذف شد.

برای آماده سازی document ها علامت NOT\_ به کلماتی که بین عبارت منفی کننده مانند not, isn’t , didn’t و اولین علامت نگارشی پس از عبارت منفی کننده، اضافه شد.

This place is not OK.

Features={ This ,place, NOT\_OK)

برای حفظ سادگی این راه و بررسی تاثیر مستقیم Naïve Bayes ، از موارد پیش پردازش دیگر مانند بردن به حالت ریشه ای و حذف برخی کلمات خودداری شده است.

در این مثال 5 کلمه ی مثبت و 5 کلمه ی منفی از بررسی این نظرات به دست آمده( کلماتی که در هر دو بخش بودند حذف شده) و اطلاعات کلمات در هر کلاس به شکل زیر است.

فرض می شود که کلماتی که در این جدول های نمونه قرار ندارند به طور کلی دیده نشده اند.

Table 1-positive label class Table 2-negative label class

|  |  |
| --- | --- |
| Word | count |
| good | 10 |
| purchased | 26 |
| excellent | 24 |
| love | 16 |
| brilliant | 4 |

|  |  |
| --- | --- |
| Word | count |
| Not\_buy | 32 |
| quit | 3 |
| repair | 13 |
| Bad | 20 |
| waste | 14 |

برای مشاهده ی روند کلی این فرایند جمله ی زیر را بررسی می کنیم

Document= the watch quit working after five weeks and it requires large amount to repair this.

Features={ the, watch ,quit, working, after, five, weeks ,and, it, requires ,large ,amount ,to ,repair, this}

برای محاسبه ی احتمال مثبت بودن این داکیومنت به شکل زیر عمل می کنیم:

P(pos|document)= P(pos)\*[P("the "│pos)\*P("watch "│pos)\*P("quit "│pos)\*P("working "│pos)\*P("after "│pos)\*P("five "│pos)\*P("weeks "│pos) \*P("and "│pos) \*P("it "│pos) \*P("requires "│pos)\*P("large "│pos)\*P("amount "│pos)\*P("to "│pos) \*P("repair "│pos) \*P("this "│pos)]

P(pos)=5/10=0.5

P("the "│pos)= P("watch "│pos)= P("after "│pos)=…= =

P(pos|document)=0.5\*=5\*10-31

و برای محاسبه ی احتمال منفی بودن این داکیومنت به شکل زیر عمل می کنیم:

P(neg |document)= P(neg)\*[P("the "│ neg)\*P("watch "│ neg)\*P("quit "│ neg)\*P("working "│ neg)\*P("after "│ neg)\*P("five "│ neg)\*P("weeks "│ neg) \*P("and "│ neg) \*P("it "│ neg) \*P("requires "│ neg)\*P("large "│ neg)\*P("amount "│ neg)\*P("to "│ neg) \*P("repair "│ neg) \*P("this "│ neg)]

P(neg)=5/10=0.5

P("the "│ neg)= P("watch "│ neg)= P("after "│ neg)=…= =

P("quit "│ neg)= =

P("repair "│ neg)= =

P(neg|document)= 0.5\*=280\*10-31

براساس این روش مقدار احتمال برچسب منفی برای این داکیومنت بیشتر است پس این عقیده منفی است.