پردیس علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

تشخیص احساس از روی متن

نگارنده

مریم رستگاری

استاد راهنما: دکتر هدیه ساجدی

پایان نامه برای دریافت درجه کارشناسی
در رشته علوم کامپیوتر

چکیده

با گسترش علم داده کاوی [۱] و کاوش متن [۲] در سال های اخیر توجهات زیادی معطوف به تحلیل احساسی متن شده است. کاربرد های تشخیص احساس در تجارت، سیاست، روانشناسی و هوش مصنوعی باعث محبوبیت هر چه بیشتر آن شده است. اصولا احساسات در انسان با عوامل مختلفی مشخص می گرد مثل حالت چهره، فشار خون، ضربان قلب و .... در این پروژه ما تمرکز خود را بر تشخیص احساس از روی متن گذاشتیم. با گسترش اینترنت و فراهم آمدن فضاهای مناسب و گسترده ی عمومی برای به اشتراک گذاشتن آزادانه نظرات و احساسات، اکنون داده های متنی زیادی در دسترس محققین است که یقینا مطالعه دستی این حجم از اطلاعات غیر ممکن است. پس به ناچار باید به دنبال روش ها و ابزار هایی بود که بتوانند به طور خودکار متن ها را به لحاظ احساسی تحلیل کنند. در این پروژه، ابزار ها و الگوریتم هایی برای حل این مساله ارائه و آن ها را به لحاظ کارایی با یکدیگر مقایسه می کنیم.

[۱] data mining
[۲] text mining

پیشگفتار

تمایل به دانستن احساسات انسان مسأله ای نیست که به تازگی ذهن بشر را به خود مشغول کرده باشد. حتی در زندگی روزمره نیز انسان علاقمند است احساس اطرافیان از اعضای خانواده و دوست گرفته تا مشتری، همکار و یا رئیس خود را بداند. هنگامی که با یک دوست در شبکه اجتماعی در حال چت هستیم، هنگامی که یک بازاریاب برای فروش محصولاتش در حال صحبت است، هنگامی که صاحبان کسب و کار های تجاری در حال خواندن نظرات مشتریان هستند، حتی هنگامی که مسئولین یک کشور در حال بررسی مطبوعات هستند، همه این ها در حقیقت خواهان اینند که بدانند مخاطبشان به چه فکر می کند و چه احساسی نسبت به آن ها دارد. از اینرو است که می گوییم این مسأله نه مسأله جدیدی است و نه حتی منحصر به حوزه خاصی است. بلکه مسأله ای است روزمره که هر کسی در هر دوره زمانی ممکن است با آن مواجه شده باشد. اما در عصری زندگی می کنیم که اینترنت بستری را فراهم کرده تا ارتباط انسان ها با هم هر چه بیشتر گسترش یابد و این، داده های بیشتری را برای بررسی و تحقیق فراهم کرده است. اکنون در هر زمینه ای، سیاسی، ورزشی، تجاری،هنری و... به راحتی می توان به نظرات زیادی دسترسی پیدا کرد. و همچنین این گستردگی داده ها محققین را به فکر واداشته تا راهی پیدا کنند که به طور خودکار بتوان این داده ها را بررسی و به عقائد و احساسات صاحبان آن ها پی برد. با وجود مطالعات زیادی که درباره این مسأله صورت گرفته، هنوز راهی که به طور قطع بتواند این مسأله را حل کند پیدا نشده است و راه هایی که ارائه می شوند تنها در صدد بهبود روش های قبلی اند. در این پروژه ابتدا چالش هایی که در حل این مساله پیش روی ما است و سپس کاربردهای آن را بیان می کنیم و در ادامه تعدادی از این پاسخ ها ارائه و با یکدیگر مقایسه شده اند و در انتها نیز یک نمونه از پیاده سازی های موفق این مسأله بیان شده است.

چالش ها

چالش های این مسأله به طور کلی به چهار دسته کلی تقسیم می شوند:
۱)جمع آوری داده: در هر روشی ما به یک مجموعه داده [۳] اولیه نیاز داریم تا بر اساس آن ویژگی هایی [۴] را به عنوان معیارهای تشخیص احساس انتخاب کنیم. جمع آوری این داده ها و برچسب دهی اولیه به آن ها کاری دشوار، و هم به لحاظ زمانی و هم به لحاظ مالی پرهزینه است.
۲)انتخاب ویژگی: انتخاب ویژگی هایی که آن ها را معیار قرار دهیم خود یکی از چالش های مسأله است و شاید بتوان گفت مهمترین قسمت حل مسأله است. اینکه چه ویژگی هایی در راه حل انتخاب شود تاثیر مستقیمی روی دقت پاسخگویی راه حل دارد.
۳)انتخاب نوع طبقه بند [۵] : گام اساسی دیگر در راه حل های این مسأله، انتخاب طبقه بند است. اینکه بهترین طبقه بند، چه طبقه بندی است و پارامتر های مربوط به آن چه باشد [۱] .
۴)ابهامات و استعاره ها: ابهامات گرامری که ذات هر زبان است ، استعاره ها و تکه کلام هایی که هر انسان ممکن است به کار ببرد کار تشخیص احساس از روی متن را بسیار دشوار می کند؛ چون از روی ظاهر کلمات و ساختار جمله، یک ماشین اتوماتیک نمی تواند به معنی حقیقی آن ها پی ببرد. حتی گاهی برای خود انسان نیز تشخیص احساس واقعی نویسنده دشوار است.
۵)این مسأله مستقل از جنسیت، فرهنگ و حتی افراد نیست چون در فرهنگ ها، جنسیت ها و انسان های مختلف یک جمله می تواند احساسات و عقاید متفاوتی را بیان کند.

کاربردها

۱)در تجارت تشخیص احساس می تواند به تحلیل بازخورد مشتری ها از محصولات و خدمات کمک کند تا صاحبان کسب و کار متوجه شوند کدام قسمت کارشان را باید تغییر دهند تا رضایت مشتری بیشتر جلب شود [۲] .
۲)تشخیص احساس می تواند به ارتباط بهتر انسان با کامپیوتر کمک کند. همچنین به سیستم

[۳] dataset
[۴] feature
[۵] classifier

های پیشنهاد دهنده کمک کند تا بر اساس حال کاربر محصول یا تعامل مناسب وی را پیشنهاد دهند(معرفی دوست در شبکه های اجتماعی یا معرفی یک کتاب مناسب) [۳] .

۳)فهمیدن احساس در هر نهاد و سازمانی و در امور مختلف می تواند مفید باشد. مثل موسسه های تجاری ، مبارزات سیاسی ، مدیریت پاسخ به یک فاجعه طبیعی و حتی ساختن ابزار های بهتر برای هوش مصنوعی [۴] .

فهرست مطالب

فصل ۱

مفاهیم مقدماتی

۱.۱ مدل های روانشناسی احساس

شرط لازم برای صحبت کردن درباره استخراج احساسات این است که ابتدا یک دید کلی درباره این که مدل ها و تئوری های احساس در روانشناسی چه هستند داشته باشیم. در این قسمت از پروژه درباره تعریف ها، ا صطلاحات، مدل ها و تئوری ها صحبت می کنیم. جامع ترین و مقبول ترین تئوری ها را به طور خلاصه بیان می کنیم تا اطلاعات مورد نیاز برای ادامه پروژه را به خواننده بدهیم.

در روانشناسی احساسات به دو دسته ساده و پیچیده(یعنی احساساتی که طبقه بندی آن ها با داشتن تنها یک عبارت دشوار است. مثل حس گناه ، شرم، غرور و ..) تقسیم می شوند. در این پروژه وقتی درباره احساس صحبت می کنیم اکثرا منظورمان احساسات ساده هستند.

اگرچه که هیچ مدل جامع قابل قبولی برای احساس وجود ندارد اما تعدادی از مدل هایی که در تشخیص احساس مورد استفاده قرار می گیرند و به طور گسترده مورد قبول هستند را به طور خلاصه بیان می کنیم. این مدل هابه دو دسته تقسیم می شوند: مدل های گسسته و مدل های ابعادی(پیوسته). با توجه به تئوری احساس گسسته بعضی احساسات بر اساس مشخصه های عصبی روانشناسی رفتاری قابل تشخیصند [۵] .

یکی از مثال های معروف و پرکاربرد احساسات شش گانه اکمن [۱] است. اکمن در یک مطالعه که بر اساس فرهنگ های متقابل بود به شش دسته احساس رسید: غم، خوشحالی، خشم، ترس، نفرت و هیجان [۶] .

اکثر مقالات در باب تشخیص احساسات از این مدل استفاده می کنند در حالیکه بعضی نیز از مدل چرخ احساس پلاتچیک [۲] استفاده می کنند. که در آن پلاتچیک هشت احساس ابتدایی را دسته بندی می کند: لذت، اعتماد، ترس، هیجان، غم، نفرت، خشم و انتظار [۷] .

پرروت [۳] در دسته بندی سه لایه خود شش حس اولیه را در نظر گرفت: عشق، لذت، هیجان، خشم، غم و ترس در لایه اول که با ۲۵ حس ثانویه دنبال می شود و در آخرین لایه نیز حس ها را با جزییات بیشتری دسته بندی کرده است [۸] .

در یک دیدگاه متفاوت مدل ابعادی سعی می کند احساسات را بر اساس دو یا سه بعد بیان کند. بر خلاف تئوری احساسات ساده که بیان می کند احساسات مختلف به زیرسیستم های عصبی متفاوت در مغز مربوطند مدل تئوری ابعادی بر اساس این فرض است که تمام احساسات بر اساس یک سیستم مشترک و بهم پیوسته نوروفیزیولوژیک است. مدل ارائه شده توسط راسل بیان می کند که احساسات را می توان در یک فضای دایروی دو بعدی نشان داد که یک بعد برای شدت و یک بعد برای میزان دلپذیری است. مدل ابعادی به ندرت در تشخیص احساس مورد استفاده قرار می گیرد [۴] .

۲.۱ یادگیری ماشین

یادگیری ماشین یکی از زیرشاخه های علم هوش مصنوعی است که معمولا از روش های آماری استفاده می کند تا حالت یادگیری را برای کامپیوتر شبیه سازی کند [۹] .

[۱] Ekman
[۲] Plutchik
[۳] Parrott

ماشین پشتیبانی بردار

ماشین های پشتیبانی بردار(Support Vector Machine (SVM)) مدل های یادگیری با نظارتی هستند که داده را برای طبقه بندی تحلیل می کنند. با داشتن مجموعه ای از داده های ورودی که هر کدام از داده های آن با یک یا چند طبقه [۴] برچسب گذاری شده اند،این ماشین مدلی می سازد تا بتواند به ورودی جدید یک یا چند طبقه نسبت دهد.

یادگیری عمیق

یادگیری عمیق از زیر شاخه های یادگیری ماشین است. این روش ویژگی را به صورت سلسله مراتبی از لایه های مختلف از طریق توابع غیر خطی استخراج می کند. ورودی هر لایه، خروجی لایه قبلی است و آموزش آن می تواند به صورت با ناظر یا بدون ناظر باشد. در واقع تک لایه مخفی در شبکه عصبی با تعداد زیادی لایه جایگزین شده است [۱۱] .

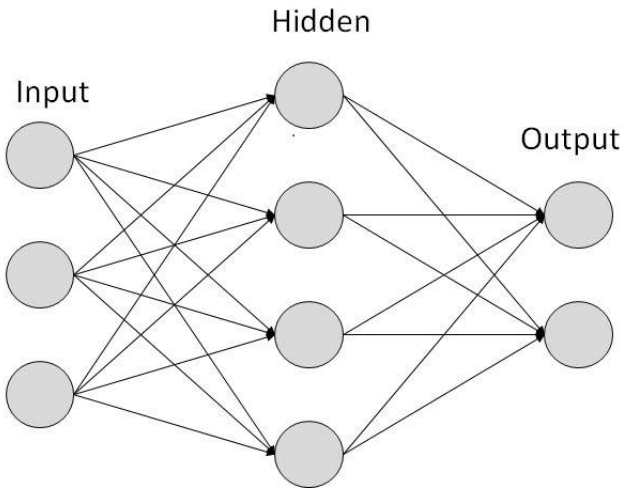
شبکه عصبی مصنوعی

یک شبکه عصبی(Artificial Neural Network) ایده ای برای پردازش است که از سیستم عصبی زیستی الهام گرفته و مانند مغز به پردازش اطلاعات می پردازد. عنصر کلیدی این ایده، ساختار جدید سیستم پردازش اطلاعات است. این سیستم از شمار زیادی عناصر پردازشی فوق العاده پیوسته به نام نورون ها تشکیل شده که برای حل یک مسأله با هم هماهنگ عمل می کنند [۱۰] .

شبکه های عصبی پیچشی

شبکه های عصبی پیچشی (Convolutional Neural Network ((CNN) یکی از مهمترین روش های یادگیری عمیق هستند که در آن ها چندین لایه با روشی قدرتمند آموزش می بینند. این روش بسیار کارآمد بوده و یکی از رایجترین روش ها در کاربرهای مختلف بینایی کامپیوتر است. یک شبکه CNN از سه لایه اصلی تشکیل می شود که عبارتند از: لایه کانولوشن [۵] ، لایه پولینگ

[۴] class
[۵] convolution



شکل۱: مثالی از یک شبکه عصبی

[۶] و لایه تماما متصل. در هر شبکه عصبی پیچشی دو مرحله برای آموزش وجود دارد. مرحله رو به جلو [۷] و مرحله پس انتشار. [۸] در مرحله اول تصویر ورودی به شبکه تغذیه می شود و این عمل چیزی جز ضرب نقطه ای بین ورودی و پارامترهای هر نورون و نهایتا اعمال عملیات کانولوشن در هر لایه نیست. سپس خروجی شبکه محاسبه می شود [۱۵] .

۱.۲.۱ شبکه عصبی بازگشتی

شبکه عصبی بازگشتی(Recurrent Neural Network (RNN)) یکی از انواع شبکه های عصبی مصنوعی است که در پردازش زبان طبیعی کاربر زیادی دارد. ایده ی اصلی این شبکه استفاده از اطلاعات متوالی است. شبکه عصبی های سنتی فرض می کنند که همه ورودی و خروجی ها مستقل از یکدیگرند. ولی این شبکه ها وقتی بخواهید کلمه بعدی در یک توالی را حدس بزنید کارا نیستند. اما شبکه های عصبی بازگشتی، حافظه دارند و اطلاعاتی که پردازش می کند را در حافظه اش نگه می دارد و در این کاربردها بسیار مناسبند [۱۲] .

[۶] pooling
[۷] feed forward
[۸] backpropagation

۳.۱ مفاهیم مقدماتی گراف

در این قسمت با فرض اینکه خواننده به مفاهیم اولیه و تعریف گراف آگاه است تنها به یادآوری بعضی از مفاهیم گراف می پردازیم.

بردار ویژه و مقدار ویژه

برداری است که وقتی روی بردار اعمال می شود، حاصل، ضرب خود بردار در عددی اسکالر می شود. به آن عدد اسکالر نیز مقدار ویژه می گویند.برای مثال در عبارت زیر ،A بردار ویژه و λ مقدار ویژه است.

$$AV = \lambda V$$

مرکزیت بردار ویژه

مرکزیت بردار ویژه، معیاری برای میزان تاثیر یک گره در یک شبکه است. مرکزیت یک راس در یک گراف به شکل زیر محاسبه می شود:

$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$

که در آن λ مقدار ویژه، $M(v)$ گره های مجاور راس V و $a_{v,t}$ مقدار ماتریس مجاورت به ازای رئوس v و t است [۱۳] .

ضریب خوشه بندی

ضریب خوشه بندی، معیاری است که مشخص می کند گره ها در یک گراف تا چه میران تمایل دارند با هم یک خوشه تشکیل دهند [۱۴] .

فصل ۲

روش شناسی

۱.۲ شیوه های نمایش مستند

یکی از اساسی ترین گام ها در راه حل هایی که برای این مسأله ارائه شده است این است که مدلی برای نمایش مستندات در الگوریتممان تعریف کنیم.

۱.۱.۲ شیوه های نمایش رایج

مدل ویژگی n-gram

این استاندارد ترین شیوه نمایش در کارهای طبقه بندی متن از جمله طبقه بندی احساسی متن است. متن ها به صورت لیستی نامرتب از عبارات و کلمات نمایش داده می شوند. با این حال محدودیت هایی نیز در این روش وجود دارد و آن این که این روش نمی تواند مفاهیم ضمنی را نمایش دهد.

مدل ویژگی n-gram خاص

این ویژگی ها برای تشخیص احساس در عباراتی که احساسات در آن ها به صورت زیرکانه ای به کار رفته طراحی شده است.
برای نمونه [۲۲] ویژگی هایی همچون positional n-gram (یعنی n-gram های نیمه اول توئیت و n-gram های نیمه دوم توئیت) و Part Of Speech (POS)، تا n-gram معمولی

را کامل کند. مشابه یافته ها برای تحلیل احساسی، positional n-gram عملکرد طبقه بندی را کاهش و POS باعث بهبودی جزئی در عملکرد شد [۳۲] .

شیوه ویژگی های لغتنامه محور

لغتنامه ای که نشان دهنده احساس هر کلمه و یا رابطه هر کلمه با هر احساس باشد، می تواند اطلاعات مفیدی برای نمایش متون در اختیار قرار دهد. [۳۳] ، به n-gram ویژگی افزود تا تعداد وقوع یک کلمه در در GPEL ها [۱] بشمارد تا عملکرد طبقه بندی وبلاگ ها را افزایش دهد. در حالیکه GPEL ها اطلاعات مفیدی درباره کلمات غنی از احساس به دست می دهند اما ایستا هستند و برای کلمات حوزه های پویا مثل توئیتر پوشش خوبی ندارند. [۲۳، ۳۴] ، نشان داده است که ویژگی هایی که بر اساس لغتنامه های DSEL [۲] هستند نسبت به لغتنامه های GPEL ، بهبود بیشتری روی n-gram ها ایجاد می کنند [۲۲] . در مجموع استفاده از دانش لغتنامه های DSEL ، که رابطه بین احساسات و کلمات را کمی می کند، می تواند برای طراحی ویژگی های بهتر برای طبقه بندی احساسی به کار رود [۳۵] .

۲.۱.۲ شیوه نمایش گراف پایه

آنچه گفتیم شیوه های نمایش مستنداتی است که تا کنون رایج بوده حال شیوه ای جدید برای این کار معرفی و عملکرد طبقه بندی با استفاده از آن را بررسی می کنیم. در این بخش یک الگوریتم استخراج ویژگی گراف پایه معرفی می کنیم که به طور خودکار مجموعه ای از الگو های نحوی غنی از احساس را استخراج می کند. ما از نماد حرف ایتالیک برای مایش اسکالر، از حروف کوچک برجسته برای نمایش بردار و حروف بزرگ برجسته برای نمایش ماتریس استفاده می کنیم. به هر یک از الگوهای $p = \{p_1, p_2, ..p_n\}$ یک وزن نسبت داده می شود که به آن امتیاز الگو می گویند که بیانگر این است که الگوی p چقدر برای احساس e اهمیت دارد. الگوها و وزن هایشان نقش مشخصه ها را برای طبقه بندها بازی می کنند. الگوریتم استخراج مشخصه گراف پایه در گام های زیر خلاصه می شود.

[۱] General Purpose Emotion Lexicon
[۲] Domain Specific Emotion Lexicon

گام اول(نرمال سازی):
ایتدا متن هایی که از طریق API Twitter به دست آمده را به دو گروه تقسیم می کنیم: گروه توییت های پربازدید که از هشتگ های پرتکرار به دست آمده اند و گروه توییت های خبری که از حساب های خبرگزاری ها استخراج شده اند. هر دو گروه مجموعه داده با توجه به فاصله توکنایز شده اند سپس به این صورت پیش پردازش شده اند: حروف بزرگ به حروف کوچک تبدیل شده اند. نظرات کاربران با <user mention> نشانی های اینترنتی با <url> و هشتگ ها با <hashtag> جایگزین می شوند.
گام دوم(ساختار گراف):
بعد از نرمال کردن توییت های پربازدید و توییت های خبری دو گراف گراف خبری $G_o(V_o; A_o)$ و گراف پربازدید $G_s(V_s; A_s)$ را به این صورت ساخته می شوند: مجموعه رئوس(V) نشاندهنده توکن هایی هستند که از بدنه اصلی متن ها به دست آمده اند. و مجموعه یال ها(A) رابطه بین کلمات را نشان می دهد. این رابطه از قطعه ای از متن(نه کل آن) با استفاده از رویکرد پنجره ای به دست می آید. با این روش ساختار نحوی جملات حفظ می شود. برای مثال پست زیر :

"<usermention> last night's concert was just awesome !!!!! <hashtag>"

به عبارت زیر تبدیل می شود:

$$< usermention > \Rightarrow last, last \Rightarrow night, ..., !!!!! \Rightarrow < hashtag >$$

گام سوم(ساختن گراف):
هدف از این گام به دست آوردن مجموعه ای از یال ها است که بیشتر به عبارات احساسی مربوطند. در اینجا فرض بر این است که با انطباق گراف G_s با گراف G_o به گراف G_e می رسیم که به آن گراف احساس نیز می گوییم. این گراف توکن هایی که به لحاظ احساسی مرتبطند را نگه می دارد. که با گام زیر به دست می آیند:
به ازای هر یال $a_i \in A$، وزن نرمال شده اش با معادله زیر محاسبه می شود.

$$w(a_i) = \frac{freq(a_i)}{max_{j \in A} freq(a_i)}$$

که $freq(a_i)$ تعداد تکرار یال a_i است.

۲)متعاقبا وزن های جدید برای یال های $a_i \in G_i$ با معادله زیر به دست می آید.

$$w(a_i) = \begin{cases} w(a_{s_i}) - w(a_{o_j}) & \text{if } a_{o_j} = a_{s_i} \in G_o \\ w(a_{s_i}) & \text{other wise} \end{cases}$$

وزن های به دست آمده گراف G_e به طوری تنظیم شده اند که پرتکرار ترین یال های G_o در G_e تضعیف شده اند. در نتیجه یال های با وزن بیشتر در G_e نمایش دهنده توکن هایی هستند که بیشتر به مجموعه داده خبری مربوطند. علاوه بر این منحنی های $a_i \in A_e$ بر اساس یک آستانه ϕ_w [۳] هرس شده اند.

گام چهارم(دسته بندی توکن ها):

اکنون ماتریس مجاورت M به صورت زیر تعریف می شود:

$$M_{i,j} = \begin{cases} ۱ & \text{if node i and j are linked in} G_o \\ ۰ & \text{other wise} \end{cases}$$

سپس مرکزیت بردار های ویژه و ضرایب خوشه بندی تمام رئوس V_e محاسبه می شوند. که برای دسته بندی توکن ها به دو نوع موضوعی و ربط دهنده استفاده می شود. کلمات ربط دهنده: برای اندازه گیری تاثیر تمام رئوس در گراف G_e از مرکزیت بردارهای ویژه استفاده می شود که به صورت زیر به دست می آید:

$$c_i = \frac{۱}{\lambda} \sum_{j \in V_e} M_{i,j} c_j$$

که λ نشاندهنده فاکتور نسبت است و c_i امتیاز مرکزیت راس i ام است. با داشتن λ به عنوان مقدار ویژه متناظر می توان این معادله را به فرم برداری $MC = \lambda C$ نوشت که C یک بردار ویژه ماتریس

[۳] این آستانه به طوری تجربی به دست می آید.

M است. با داشتن یک بردار ویژه انتخاب شده ی C و امتیاز مرکزیت بردار ویژه راس i ام که با c_i نشان می دهیم، لیست نهایی کلمات ربط دهنده که در ادامه آن را با CW نشان می دهیم با نگه داشتن توکن های $c_i > \phi_{eig}$ [۴] به دست می آید.

این لیست نشاندهنده کلماتی است که پرتکرارند و مرکزیت بالایی دارند(مثل "or" و "and" و "my") .

۲)کلمات موضوعی: این کلمات معمولا با هم خوشه بندی می شوند. یعنی کلمات موضوعی زیادی با کلمات ربط دهنده مشابه به هم متصلند. بنابر این یک ضریب به همه رئوس G_e نسبت داده می شود که به شکل زیر محاسبه می شود.

$$cl_i = \frac{\sum_{j\neq i;k\neq j;k\neq i} M_{i,j} \times M_{i,k} \times M_{j,k}}{\sum_{j\neq i;k\neq j;k\neq i} M_{i,j} \times M_{i,k}} \times \frac{۱}{|V_e|}$$

که در آن cl_i میانگین ضریب خوشه بندی راس i است. که میزان به هم متصل بودن همسایه های راس i را نشان می دهد. مشابه کلمات ربط دهنده که از این به بعد کلمات موضوعی را SW با نشان می دهیم که با نگه داشتن توکن های $cl_i > \phi_{cl}$ [۵] به دست می آید

(برای مثال کلماتی چون "never" و "life") .

گام پنجم(الگو های داوطلب):

با داشتن مجموعه توکن های SW و CW الگو های کاندید با رویکردی خودکار ساخته می شود به طوریکه که ساختار نحوی را نیز از دست ندهیم. در ادامه این قوانین را برای تعریف الگوهای داوطلب می آوریم: <sw،sw،cw> و <sw،cw،sw> و <cw،sw،sw> و <cw،cw،sw>. که sw و cw نشاندهنده توکن های دلخواه از مجموعه SW و CW هستند. مهم است که روشن کنیم در این کار از توالی های به طول دو و سه استفاده شده است چون به طور تجربی نتیجه گرفته شده که برای کار ما مناسب تر هستند. ممکن است در ادامه گاهی اوقات به جای الگوی داوطلب از لفظ قالب [۶] استفاده کنیم. تفاوت این کار با کارهای مشابه در این است که از قوانین و ابتکارات گرامری در فرآیند استخراج الگوها استفاده نشده است برای همین این الگوهاییهتر عبارات صریح و هم ضمنی را پوشش می دهند.

[۴] این آستانه به طوری تجربی به دست می آید.
[۵] ϕ_{cl} به طور تجربی به دست می آید.
[۶] template

جدول۱ : مثالی از الگوها و قالب های استخراج شده در فرآیند استخراج الگوی ساده

Templates	Examples Pattern
<cw،sw>	“stupid*”،“like*”،“am*”
<cw،cw،sw>	“shut up*” ، “love you*”
<sw،cw،sw>	“*for*”
<sw،cw،cw>	“* on the”
<sw،cw>	“*<hashtag>”

۶)گام ششم(استخراج الگو): یک فرآیند استخراج الگوی ساده شامل به کار بردن قالب های نحوی روی متن های آموزشی می شود. علاوه بر این هر کلمه موضوعی sw در هر الگو با یک <*> جایگزین می شود. این کار باعث می شود که وقتی رو یک متن که در متون آموزشی نیست مدلسازی می شود کلمات موضوعی که در داده آموزش نبوده را نیز در نظر بگیرد. ما به الگوهایی علاقمندیم که ارتباط بیشتری با موضوعیت دارند بنابراین الگوهایی که بیشتر از یک آستانه تکرار می شوند را نگه میداریم و بقیه را دور می ریزیم. در جدول۱ مثالی از الگوها و قالب های متناظرشان که در استخراج الگوی ابتدایی استخراج شده اند آورده شده است.
در ادامه در باره فرآیند گسترش الگوهای نحوی با کلمات جاساز شده صحبت می کنیم. این فرآیند گسترش کمک می کند معنای بین الگوها حفظ شود و ارتباط ویژگی ها را بهبود می دهد.

الگوهای گسترش یافته

کلمات جاسازی شده وزن دار: ابتدا کلمات جاسازی شده که براساس توییتر از پیش آموزش دیده شده اند را از [۱۶] به دست آورده شده اند. و بر اساس متون احساسی دوباره وزن آن ها را محاسبه شده اند [۱۷] . سپس یک شبکه عصبی عمیق کاملا متصل با یک لایه پنهان را طی ده دوره با استفاده از پس انتشار آموزش [۷] داده شده اند. همانطور که در [۱۶] آورده شده است. کلمات احساسی جاسازی شده را با $W \in \mathbb{R}^{d \times n}$ نشان می دهیم که $d = ۵۲$ است. توجه شود که از tf-idf برای کاهش تعداد کلمات لغتنامه استفاده شده است.(از ۱۴۰ K به ۲۰ K کلمه)
خوشه های کلمات: سپس با استفاده از اطلاعات کلمات جاسازی شده و با روش تجمعی کلماتی که به لحاظ معنایی مرتبطند را خوشه بندی می کنند. برای مقایسه کیفیت خوشه بندی آن را با

[۷] backpropagation

جدول۲: مثالی از یک لغتنامه WordNet-Affect

Emotion	Words
Anger	irascibility, short-temper, spleen
Fear	frighten, fright, scare
Joy	hilarious, screaming, uproarious
Sadness	penitently, penitentially
Surprise	astonishing, astounding, staggering
Disgust	detestably, repulsively, abominably

WordNet-Affect [۸] مقایسه می کنیم. از روش Ward [۱۸] به عنوان ملاک اتصال و از فاصله کوسینوسی به عنوان معیار فاصله ای استفاده شده است برای پیاده سازی خوشه بندی از scikit-learn [۱۹] استفاده شده است.

ساختن الگوی گسترش یافته: هدف خوشه های کلمات این است که از آنها را استفاده شود تا فرآیند گسترش الگو هدایت شود.به عبارت دیگر، الگوها تعدادی رابطه معنایی را نگه می دارند که برای طبقه بندی مفید است. توجه کنید که فرآیند شبیه استخراج الگوی ساده است به جز اینکه در این روش از کلمات جاسازی شده نیز استفاده می کند. این روش مستلزم یک فرآیند خودکار است که یک متن را پردازش کند و در یک روش جامع الگوهای داوطلب جستجو کند. هر توالی از کلمات متون احساسی را که در شرایط قالب ها بگنجد را نگه می داریم و بقیه را دور می ریزیم. به علاوه مولفه ی sw قالب ها باید کلمه ای باشد که در خوشه های بالا یافت شود. الگوهایی که کمتر از ده بار تکرار می شوند حذف می شوند و در مجموع ۱۸۷۶۴۷ الگو تولید می شود.

وزن دهی الگوی احساس

الگوهای به دست آمده در گام های پیشین هنوز به هیچ دسته احساسی نسبت داده نشده اند. قبل از آموزش یک مدل طبقه بندی یک مکانیزم وزن دهی الگو باید به کار گرفته شود. همانند سایر مکانیزم های محبوب وزن دهی مثل tf-idf وزن مشخص کننده اهمیت الگوها برای هر احساس است. چنین وزن دهی از تغییر tf-idf به دست می آید. که آن را emotion frequency(pf-ief)

[۸] مجموعه ای از کلمات(شامل اسم صفت قید و فعل) و احساسات مربوط به هر کلمه است که به طور دستی جمع آوری شده و به چهار دسته مثبت، منفی، خنثی و مبهم و بیست و هشت زیردسته لذت، عشق، ترس و ...تقسیم شده است.

pattern frequency-inverse می نامیم و در دو گام تعریف می کنیم. ابتدا pf به صورت زیر محاسبه می شود:

$$pf_{p,e} = \log \frac{\sum_{p_i \in P_e} freq(p_i, e) + ۱}{freq(p_i, e) + ۱}$$

که $freq(p, e)$ ،تکرار p در e و $pf_{p,e}$ لگاریتم تکرار الگوی p در مجموعه ای از متون مرتبط با احساس e است. حال ief به صورت زیر محاسبه می شود:

$$ief_p = \log \frac{freq(p_i, e) + ۱}{\sum_{e_j \in E} freq(p, e_j) + ۱}$$

که ief_p معیار ارتباط الگوی p با تمام دسته های احساسات است. در نهایت امتیاز الگو را به صورت زیر تعریف می کنیم:

$$ps_{p,e} = pf_{p,e} \times ief_p$$

که $ps_{p,e}$ نشاندهنده این است که الگوی p تا چه حد برای احساس e مهم است.

۲.۲ مدل های تشخیص متن

DeepEmo ۱.۲.۲

این مدل یک شبکه عصبی چندلایه را با شکل ماتریسی ویژگی های گراف محور ترکیب می کند. ورودی شبکه یک ماتریس m در n است $(X \in \mathbb{R}^{m \times n})$ که مولفه های آن بیانگر امتیاز الگوی گسترش یافته i در احساس j هستند. روی خروجی این فرآیند یک تابع فعالسازی غیر خطی اعمال می شود و یک ماتریس نگاشت ویژگی تولید می کند. یک ادغام روی هر نگاشت ویژگی اعمال می شود. نتایج ادغام با همان ترتیب به عنوان ورودی به دو لایه پنهان با اندازه های ۵۱۲ و ۱۲۸ داده می شود سپس یک دسته ۱۲۸تایی از آنها انتخاب و توسط بهینه ساز Adam در هفت دور آموزش می بیند. در نهایت از یک تابع برای ساخت طبقه بندی نهایی استفاده شده است.(ما برای پیاده سازی شبکه عصبی از Keras استفاده شده است.) [۲۰]

۲.۲.۲ مدل برداری

در این قسمت یک مدل برداری ساده(EVM) ارائه می شود. که استفاده و کاربرد الگوی پیشنهاد شده قسمت ۲.۱.۲ را نشان می دهد. وزن الگوها از مکانیزیم معرفی شده بخش وزن دهی الگوی احساس به دست می آید. اگر n الگو و m احساس داشته باشیم کل مدل احساسی را با یک ماتریس $EM \in \mathbb{R}^{n \times m}$ نشان می دهیم. که مولفه های $EM_{i,j}$ رتبه الگوی i در احساس j هستند. که بر اساس امتیاز الگوی $ps_{i,j}$ تعیین می شود. هر چه الگو به آن احساس مرتبط تر باشد ps مقدار بیشتر و رتبه کمتری دارد. فرض کنید یک پست اجتماعی tw داریم. ابتدا بردار تکرار $f \in \mathbb{R}^n$ آن به دست آورده می شود. که f_i تکرار الگوی i در پست ورودی d را نشان می دهد. امتیاز احساس را به صورت زیر محاسبه می شود:

$$es = f.EM$$

که $es \in \mathbb{R}^m$ و هر مولفه ی $es_{i,j}$ امتیاز نهایی احساس j در پست tw است. ایندکس کمینه این مقادیر به عنوان احساس پست انتخاب می شود.

۳.۲ مقایسه مدل ها

۱.۳.۲ مدل های سنتی

در این قسمت روش DeepEmo با روش های سنتی(مثل Bag of words ، TF-IDF ، n-grams) رایج در طبقه بندی جملات مقایسه می شود.طبقه بندی که برای آموزش این مدل ها استفاده شده طبقه بند stochastic gradient descent (SGD) تهیه شده توسط scikit-learn است.

۲.۳.۲ مدل های یادگیری عمیق

معماری یادگیری عمیق ما را قادر به یادگیری خودکار مشخصه ها از اطلاعات متنی می کند. روش های یادگیری عمیق با انتخاب ورودی های متفاوت از هم متمایز می شوند، تفاوت این کار در استفاده از نمایش گرافی گسترش یافته به عنوان ورودی است. در این بخش مدل پیشنهاد شده با

جدول۳: مقایسه روش DeepEmo با سایر روش ها

Models	Features	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	F1 Avg.
BoW	**word frequency**	0.53	0.08	0.17	0.53	0.71	0.60	0.36	0.33	0.57
BoW_{TF-IDF}	**TF-IDF**	0.55	0.09	0.18	0.57	0.73	0.62	0.39	0.35	0.60
n-gram	**word frequency**	0.56	0.09	0.17	0.57	0.73	0.64	0.42	0.39	0.61
$n\text{-}gram_{TF-IDF}$	**TF-IDF**	0.58	0.12	0.17	0.60	**0.75**	0.67	0.47	0.45	0.63
char	**character frequency**	0.35	0.03	0.04	0.20	0.51	0.46	0.10	0.12	0.37
$char_{TF-IDF}$	**TF-IDF**	0.33	0.03	0.06	0.21	0.52	0.45	0.11	0.13	0.37
char_ngram	**character frequency**	0.49	0.06	0.12	0.46	0.67	0.55	0.30	0.28	0.52
$char\_ngram_{TF-IDF}$	**TF-IDF**	0.53	0.07	0.15	0.53	0.71	0.59	0.35	0.31	0.57
word2vec	**word embeddings**	0.50	0.02	0.13	0.48	0.69	0.51	0.35	0.31	0.53
LIWC	**affect words**	0.35	0.03	0.11	0.30	0.49	0.35	0.18	0.19	0.35
EVM	**patterns**	0.42	0.02	0.04	0.38	0.50	0.34	0.24	0.21	0.38
CNN-patt	**basic patterns**	0.47	0.00	0.00	0.45	0.67	0.61	0.15	0.08	0.52
DeepEmo	**enriched patterns**	**0.58**	**0.16**	**0.32**	**0.65**	**0.75**	**0.70**	**0.59**	**0.55**	**0.67**

جدول۴:توزیع احساسات در مجموعه داده های آموزش و آزمون

Emotions	Train	Test	Hashtags
sadness	192842	21422	#depressed, #grief
joy	149986	16663	#fun, #joy
fear	92145	10209	#fear, #worried
anger	91947	10200	#mad, #pissed
surprise	41337	4691	#strange, #surprise
trust	17295	1913	#hope, #secure
disgust	8052	873	#awful, #eww
anticipation	3588	384	#pumped, #ready

مدل های شبکه های عصبی پیچشی(CNNs) ، شبکه های عصبی بازگشتی(RNNs)، شبکه های عصبی با گیت های دو طرفه (GRNNs) مقایسه می شود.

۴.۲ آزمایش ها

۱.۴.۲ داده ها

ابتدا مجموعه ای از هشتگ ها ساخته می شود تا توییت های انگلیسی از API Twitter به دست آورده شود. این کار هشت حس ابتدایی پلاتچیک را مبنای کار خود قرار می وهد: خشم، انتظار، نفرت، ترس، لذت، غمگینی، هیجان و اعتماد.
در مجموع ۳۳۹ هشتگ تعریف شد. برای اینکه از کیفیت هشتگ ها مطمئن شویم از گام های پیش پردازش پیشنهاد شده ی [۲۱، ۲۲، ۲۳] استفاده شده است و فرض شده است هشتگ در

جدول۵:نتایج آزمایش ها با سایر روش های یادگیری

Model	Adopted from	Input	Epochs	Accuracy
RNN	(ZOLKEPLI, 2017)	word2vec (Mikolov et al., 2013)	24	0.53
CNN	(Kim, 2014)	character embeddings (end-to-end)	50	0.63
Bi-GRNN	(Ivanov, 2017)	enriched patterns (ours)	12	**0.65**

آخرین مکان یک توییت ظاهر می شود. داده ها به داده ی آموزش و آزمون به نسبت ۹ به ۱ تقسیم شده است. در قسمت های بعدی کارایی الگو های گسترش یافته در کارهای تشخیص احساس متفاوت تکامل داده شده است. از امتیاز بهF۱ عنوان معیار تکامل استفاده شده است.

۲.۴.۲ نتایج تجربی

نتایج با استخراج کننده های سنتی

نتایج به دست آمده از استخراج کننده های سنتی در جدول شماره۳ آمده است. همانطور که جدول نشان می دهد مدل های TF-IDF معمولا نتایج بهتری در مقایسه با ویژگی هایی که بر اساس شمارش اند در هر دو سطح حرف و کلمه استخراج مشخصه به دست می دهد.

نتایج با رویکرد های الگویی

نتایج EVM و CNN-patt که الگوهای گراف محور را به کار می گیرند از اکثر روش های متداول بدترند. DeepEmo که از الگوهای گسترش یافته استفاده می کند نتایج بهتری نسبت به هر دو EVM و CNN-patt و همه روش های رایج به دست می آورد. (امتیاز F۱ آن برابر ۶۷ درصد است.) در حقیقت روش DeepEmo بهترین امتیاز F۱ را به ازای همه احساس ها به دست می آورد. در مقایسه با مدل های الگوهای ابتدایی(CNN-patt) وقتی از الگوهای گسترش یافته استفاده می کنیم شاهد بهبود معناداری در عملکرد هستیم(مثبت ۱۵ درصد). به طور کلی استفاده از ویژگی های گراف محور گسترش یافته برای آموزش مدل های تشخیص احساس امکان پذیر و کارا است.

جدول۶:مقایسه روش ما با سایر روش ها

Method	Data Size	F1-score
Roberts (2012)	3777	0.67
Qadir (2013)	4500	0.53
Mohammad (2015)	21,051	0.49
Volvoka (2016)	52,925	**0.78**
DeepEmo (Ours)	597,192	**0.72**

مقایسه با آخرین دستاوردها

همچنین نتایج به دست آمده این روش با سیستم‌هایی که بر اساس احساسات شش گانه اکمن هستند نیز مقایسه شده اند. برای مقایسه عادلانه مجموعه داده ها از هشت احساس به شش احساس خشم، نفرت،ترس ،لذت،غمگینی و هیجان کاهش داده شده اند. همانطور که در جدول۶ نشان داده شده، سیستم تشخیص احساس پیشنهاد داده شده به نسبت همه روش ها به جز (Bachrach، ۲۰۱۶ Volvoka and) [۲۴] نتایج بهتری کسب می کند.(امتیاز F۱ آن ۷۲ درصد است.).سیستم آنها بهتر از سیستم پیشنهادی عمل می کند(امتیاز F۱ آن ۷۸ درصد است) به خاطر اینکه از مشخصه های خوش تعریف زبانی مثل هشتگ استفاده می کند.مشخصه های سیستم پیشنهادی به نویز حساس ترند برای اینکه هدف پوشش بیشتر عبارات احساسی ضمنی بوده است. علاوه براین ویژگی های آنها دامنه معین دارند که یعنی بعضی از ویژگی های مهم آنها مثل هشتگ ها ممکن است برای مجموعه داده های موثر دیگر کاربردی نباشند.
با توجه به [۲۵] روش های سنتی روی این دسته از وظایف برای مجموعه داده ها با اندازه تا چند صد هزار قدرتمند ترند و با بزرگتر شدن مجموعه داده هنگامی که اندازه آن چند میلیونی می شود مدل های CNN شروع به بهتر عمل کردن می کنند.

نتایج با یادگیری عمیق

در این قسمت یک مقایسه بین مدل های مختلف یادگیری عمیق متفاوت که بر اساس احساسات شش گانه اکمن هستند را پیشنهاد می شود. الگوهای گسترش یافته به عنوان ورودی به یک GRNN داده می شوند و همانطور که در جدول شماره۵ می بینید بهترین نتایج بین مدل های مختلف یادگیری عمیق به دست می آید(دقت ۶۷ درصد).

جدول۷:مثالی از هزار الگو در مرد و زن که بیشترین تکرار را داشتند.

Emotions	Male patterns	Female patterns
Anger	a{crazy}, you{despise}, like{try}	my{yelling}, would {want}, hate {you}
Sadness	your {lyrics}, {bouncing} your	better {come}, you {wreck}, {despise} going
Surprise	last {second}, to {announce}	happy {birthday}, {only} person
Fear	{you} have, {getting} dark	my {stepmom}, the {loneliest}

۵.۲ تحلیل الگوهای گسترش یافته

در این قسمت الگوهای گسترش یافته‌ای که از مجموعه داده جنسیت محور استخراج شده بررسی می شود. ابتدا از توییتر مجموعه داده جمع آوری شده است و کاربران بر اساس محتوا توسط پیش بینی کننده جنسیت [۳۶] به مرد و زن طبقه بندی شده اند. با این کار یک مجموعه داده جنسیتی تولید شده اند و صورت دستی، درستی آن را تایید شده است. به طور تصادفی ۲۰۰۰ مرد و ۲۰۰۰ زن نمونه برداری شده اند. در مجموع این کار ۴۰۰۰۰۰ توییت را تولید می کند که توییت ها با کمتر از ۵ کلمه دور ریخته می شود.نهایتا توییت ۲۹۴،۷۹۲ خواهیم داشت.که با استفاده از DeepEmo آنها را طبقه بندی شده اند. سپس این الگوها تحلیل شده اند و الگوهای مشترک بین مردان و زنان را دور ریخته شده و هزار الگوی برتر در هر جنسیت را تحلیل می شوند. مثال هایی از پرتکرارترین الگوهای احساسی به دست آمده از قالب های <cw،sw> و <sw،cw> در جدول شماره۷ نشان داده شده است. کلمات داخل نشاندهنده کلمات موضوعی اند که در فرآیند گسترش الگو به دست آمده اند. همانطور که مشاهده می کنیم کلمات موضوع نشاندهنده کلمات غنی از احساس مثل despise وyelling و loneliest هستند. کلمات متصل کننده مفهوم را منتقل می کنند که کمک می کنند تا الگوهای گسترش یافته را بهتر بفهمیم.

اکنون این پرسش را مطرح می شود که آیا عبارات و الگوهای احساسی جنسیت محور در شبکه اجتماعی وجود دارد؟ البته خیلی زود است که با تحلیل های اولیه ای که بیان شد بخواهیم نتیجه بگیریم.اما تحقیقات جالب دیگری می تواند برای پیش بینی مستقیم جنسیت انجام شود.تا اینجا هدف تحلیل بررسی الگوهای گسترش یافته بود و اینکه نشان دهیم چقدر می توانند برای تحلیل دقیقتر متون احساسی مورد استفاده قرار بگیرند.

جدول۸ میزان پوشش دهی کلمات الگوهای گسترش یافته در مجموعه داده های متفاوت. عدد داخل پرانتز نشاندهنده تعداد طبقه های آن مجموعه داده است.

Emotion Dataset	Study	Task	Domain	Dataset Size	Enriched Pattern
Our Full Dataset	Ours	Emotion (8)	Tweets	1,896,849	0.94
Gender Data	Ours	Emotion (8)	Tweets	294,792	0.89
SemEval07 Task 14	(Strapparava and Mihalcea, 2007)	Emotion (3)	Headlines	601	0.62
SemEval17 Task 4	(Rosenthal et al., 2017)	Sentiment (3)	Tweets	20,621	0.99
SemEval18 Task 1	(Mohammad and Bravo-Marquez, 2017)	Emotions (4)*	Tweets	3890	0.92
SST-2	(Socher et al., 2013)	Sentiment (5)	Reviews	58,990	0.76
SST-5	(Socher et al., 2013)	Sentiment (5)	Reviews	96,660	0.71
PsychExp	(Wallbott and Scherer, 1988)	Emotion (5)	Experiences	7339	0.95

پوشش الگو

در این قسمت پوشش الگوهای گسترش یافته روی تعدادی مجموعه داده موثر بررسی می شوند. همانطور که در جدول شماره۷ نشان داده شده ۸۹.۴ درصد توییت ها در داده جنسی شامل حداقل یکی از الگوهای گسترش یافته است.همچنین مشاهده می کنیم که اندازه مجموعه داده تاثیری روی نتایج پوشش ندارد. یک پوشش بالای ۹۵ درصدی در تجربیات احساسی [۲۶] به دست آمده که از دامنه ای متفاوت از آنچه با آن الگو ساخته شده اند نشات می گیرد. این نشاندهنده این است که الگوهای گسترش یافته پیشنهاد شده با سایر دامنه ها نیز انطباق پذیر است که فرصت هایی را برای بررسی ها و تحقیقات بعدی فراهم می کند [۲۷] .

فصل ۳

پیاده سازی

در این قسمت ما با ابزار Emo Tex آشنا می شویم.این ابزار توسط ۹k پرسش و پاسخ و نظر در تعاملات برخط آموزش و تست شده است. در ادامه شواهد تجربی را درباره عملکرد این ابزار بیان خواهد شد.Emo Tex اولین ابزار متن بازی [۱] است که هم تشخیص احساس از روی متن و هم آموزش از روی مدل های دلخواه طبقه بندی احساس را پشتیبانی می کند.

۱.۳ توصیف سیستم

سیستم با جاوا توسعه و تحت لیسانس متن باز MIT توزیع شده است. به همراه ابزار، مدل های طبقه بندی که روی مجموعه داده آن آموزش دیده است نیز منتشر شده است.

خروجی آن به این صورت است که هر متن و احساسش در یک خط بیان می شوند که در آن خط شناساگر یکتای متن و احساس مربوط به آن قرار دارد و خروجی یک فایل CSV است.(comma seprated value)

هم چنین Emo Tex می تواند برای یادگیری یک طبقه بند جدید به روی مجموعه دادهی استاندارد با برچسب های دلخواه مورد استفاده قرار گیرد. برای آموزش یک طبقه بند احساسی به مجموعه ای از متون که احساسات آن ها به طور دستی تعیین گردیده نیاز است. رویکرد یادگیری که Emo Tex پیاده کرده مستقل از چهارچوب طبقه بندیاست و بر اساس داده های آموزشی اش یاد می گیرد

[۱] open source

چگونه وجود یا عدم وجود یک حس را تشخیص دهد. برای مشاهده جزئیات بیشتر چگونگی استفاده از Emo Tex به https://goo.gl/Mjd۶y۲ مراجعه شود. سیستم بر اساس رویکردی است که در [۲۸] بیان شده کار می کند.

قبل از اینکه اقدام به استخراج مشخصه ها شود ابتدا متن ها با استفاده از کتابخانه NLP stanford [۲] توکن توکن می شوند. برچسب های html بخش های کد و نشانی های اینترنتی را که ممکن است در داده های آموزشی نویز ایجاد کنند را پاک می شوند. برخلاف [۲۸] هیچ ریشه یابی یا بن واژه سازی انجام نشده است چون ممکن است شکل اصلی کلمات حاوی اطلاعات مهمی درباره احساس آن ها باشد.

۲.۳ ارزیابی

داده ها

در این مطالعه از دو مجموعه داده استاندارد که با چهارچوب گسسته [۲۹] منطبقند، استفاده شده اند. این چهارچوب یک ساختار درختی سلسله مراتبی برای طبقه بندی احساسات تعریف می کند که در سطح بالایی خود شش احساس عشق، لذت، خشم، غمگینی،ترس و هیجان را شامل می شود.

برای آموزش طبقه بند مجموعه داده ای از ۴۸۰۰ پست که شامل پرسش و پاسخ و نظرات کاربران سایت Stack over flow است تشکیل داده شده است. این سایت مکانی برای تعامل برقرار کردن هفت میلیون برنامه نویس است که در آنجا سوالات خود را مطرح می کنند و به سوالات یکدیگر پاسخ می دهند. این نمونه گیری از میان پست های جولای ۲۰۰۸ تا سپتامبر۲۰۱۵ انجام شده است. از برنامه نویسان درخواست شد تا احساس خود را بر اساس احساسات شش گانه Shaver [۲۹] توصیف کنند. همچنین عملکرد Emo tex بر مجموعه دادهی شامل ۴۰۰۰ نظر توسعه دهندگان نرم افزار که در Jira ـ یکی از محبوبترین سیستم های پیگیری مسائل بین شرکت های نرم افزاری ـ بیان شده بود نیز بررسی شده است. مجموعه داده به دست آمده شامل احساسات عشق، لذت، خشم و ترس است. با استفاده از R package [۳۰] ،دیتا ست ها به دو قسمت ۷۰ درصد

[۲] https://stanfordnlp.github.io/CoreNLP/download.html

جدول۱: توزیع احساسات در دو مجموعه داده

Dataset	Texts conveying the emotion						N
	Love	*Joy*	*Surprise*	*Anger*	*Sadness*	*Fear*	
SO	1220	491	45	882	230	106	4800
Jira	166	124	*NA*	324	302	*NA*	4000

برای آموزش و ۳۰ درصد برای آزمون تقسیم شده اند. و از مجموعه داده آموزشی برای بهینه کردن پارامتر طبقه بندمان استفاده شده است.به ازای هر مجموعه داده مدل روی مجموعه داده آموزشی، آموزش داده و با مجموعه داده آزمون آن را ارزیابی شده است.

۳.۳ مدل آموزشی و عملکرد طبقه بند

مدل طبقه بندی Emo tex با ماشین پشتیبان بردار(SVM) [۳] آموزش داده شده است. یک SVM خطی، یک تکنیک یادگیری برای مجموعه داده های بزرگ و پراکنده ای [۴] که هم آیتم و هم ویژگی های زیادی دارند است. مثلا N، ویژگی دارند در حالیکه هر آیتم، s ویژگی غیرخالی دارد و $s \ll N$. عملکرد یک طبقه بند تا حد زیادی به پارامترهای ورودی آن بستگی دارد. تنها پارامتر در طبقه بندی خطی، پارامتر هزینه [۵] است که آن را با C نمایش می دهیم. اگر C خیلی بزرگ باشد باعث می شود هزینه طبقه بندی اشتباه زیاد شود و الگوریتم را مجبور می کند تا خطای خود را کاهش دهد ولی این می تواند منجر به بیش برازش [۶] شود. برای اینکه به پارامتر مناسبی برای SVM برسیم که هم خطای آن کمترین حالت ممکن باشد و هم دچار بیش برازش نشود، ابزار parameter tuning Liblinear روی مجموعه داده آموزشی اجرا می شود.با (۱۰-fold cross-validation) به ازای C های مختلف ۰.۰۱, ۰.۰۵, ۰.۱۰, ۰.۲۰, ۰.۲۵, ۰.۵۰, ۱, ۲, ۴, ۸ مقدار دقت محاسبه می شود و آن C که منجر به بیشترین دقت شود انتخاب می شود. نتایج عملکرد روی داده آزمون در جدول۲ آورده شده است [۳۱] .

[۳] Support Vector Machine
[۴] sparse
[۵] cost
[۶] over fitting

جدول ۲: عملکرد طبقه بند

	Stack Overflow			Jira		
Emotion	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Joy	0.77	0.77	0.77	0.85	0.85	0.85
Love	0.92	0.92	0.92	0.86	0.86	0.86
Sadness	0.79	0.79	0.79	0.83	0.83	0.83
Anger	0.86	0.86	0.86	0.75	0.74	0.74
Surprise	0.58	0.58	0.58			
Fear	0.86	0.86	0.86			

فصل ۴

نتیجه گیری

ما در این پروژه سعی کردیم روش جدیدی برای حل مسأله تشخیص احساس از روی متن معرفی و آن را با سایر روش های معمول مقایسه کنیم. برای اینکه به یک شیوه نمایش مناسب برسیم یک مکانیزم استخراج ویژگی گراف محور پیشنهاد کردیم. الگوها با کلمات جاسازی شده گسترش پیدا کردند و برای آموزش تعدادی مدل تشخیص احساس موثر به کار برده شدند. در ضمن این الگوهایعبارات دارای احساسات ضمنی را نیز دربر گرفت و نتایج تشخیص احساس را بهبود داد. در بخش انتهایی پروژه نیز، ابزار متن باز Emo Tex را برای تشخیص احساس از روی متن معرفی کردیم و دلایل تجربی بر اینکه Emo Tex ، عملکرد بسیار مناسبی روی مجموعه داده های متفاوت دارد نیز ارائه شد.

کتاب‌نامه

[1] V.Ramalingam ,A.Pandian, A.Jaiswal,N.Bhatia.Emotion detection from text.Journal of Physics:Conference Series 1000(2018)

[2] N.Gupta,M.Gilbert,G.Di Fabbrizo.Emotion detection in email customer care.Computational Intelligence 29(2013)489-505

[3] S.Voeffray.Emotion-sensitive human-computer interaction:State of the art-seminar paper.Emotion Recognition(2011)1-4

[4] A.Seyeditabar,N.Tabari,W.adrozny.Emotion Detection in Text:a Review.arXiv.org 1806.00674(2018)

[5] G.Colombetti.From affect programs to dynamical discrete emotions. Philosoph- ical Psychology 22 (2009)407-425

[6] P.Ekman.An argument for basic emotions. Cognition and emotion 6(1992)169–200

[7] R.Plutchik.Emotions: A general psychoevolutionary theory. Approaches to emotion(1984)197–219

[8] W.Gerrod Parrott. Emotions in social psychology: Essential readings. Psychology Press(2011)

[9] $http://en.m.wikipedia.org/wiki/Machine_{l}earning$

[10] $http://www.enline.ir/artificial-intelligence$

[11] $http://www.7khatcode.com/7601$

[12] http://www.wildml.com/2015/09/recurrent-neural-network-tutorial-part-1-introduction-to-rnns/

[13] $http://en.m.wikipedia.org/wiki/Eignvector_{c}entrality$

[14] $http://fa.m.wikipedia.org/wiki/$

[۱۵] آموزش شبکه های عصبی کانولوشن.سید حسین حسن پور

[16] J.Deriu, A.Lucchi, V De Luca, A.Severyn, S M¨uller, M Cieliebak, T Hofmann, M Jaggi.Leveraging large amounts of weakly supervised data for multi-language sentiment classification.Proceedings of the 26th International Conference on World Wide Web(2017)1045–1052

[17] J.Read.Using emoticons to reduce dependency in machine learning techniques for sentiment classification.Proceedings of the ACL student research workshop(2015)43-48

[18] H Ward Jr.Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301)(1963)236–244

[19] $(http://scikit-learn.org)$

[20] F.Chollet et al 2015. Keras. https://github.com/keras-team/keras

[21] M.Abdul-Mageed, L.Ungar.Fine-grained emotion detection with gated recurrent neural networks.Proceedings of the 55th Annual Meeting of the Association for Computational 1(2017) 718–728 Linguistics.

[22] W.Wang, L.Chen, K.Thirunarayan, A.P Sheth.Harnessing twitter"big data" for automatic emotion. identification.Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom).IEEE(2012) 587–592.

[23] S.M Mohammad.Emotional tweets.Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation(2012) 246–255.

[24] S.Volkova, Y.Bachrach.Inferring perceived demographics from user emotional tone and user-environment emotional contrast.Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics volume 1(2016) 1567–1578.

[25] X.Zhang, J.Zhao, Y.LeCun.Character-level convolutional networks for text classification. Advances in neural information processing systems (2015) 649–657.

[26] H.G Wallbott, R Scherer.How universal and specific is emotional experience?: Evidence from 27 countries on five continents (1998).

[27] E.Saravia, H.Toby Liu, Y Chen.DeepEmo: Learning and Enriching Pattern-Based Emotion Representations.arXive(2018)

[28] M. Ortu, B. Adams, G. Destefanis, P. Tourani, M. Marchesi, and R. Tonelli. Are bullies more productive?: an empirical study of affectiveness vs. issue fixing time. In Proc. of MSR '15. IEEE Press(2015)303-313.

[29] P. Shaver, J. Schwartz, D. Kirson, C. O'Connor. "Emotion knowledge:Further exploration of a prototype approach. Journal of Personality and Social Psychology 52(6)(1987):1061–1086.

[30] M. Kuhn et al.Caret: Classification and Regression Training. R package v. 6.0-70. $https : //CRAN.R - project.org/package = caret$, 2016.

[31] F.Calefato, F.Lanubile,N.Novielli. EmoTxt: A Toolkit for Emotion Recognition from Text.Proc.7th Affective and intelligence(2018)

[32] B. Pang , L. Lee , S. Vaithyanathan , Thumbs up?: Sentiment classification using machine learning techniques.ACL-02 Conference on Empirical Methods in Natural Language Processing, 10(2002) .

[33] S. Aman , S. Szpakowicz , Using roget's thesaurus for fine-grained emotion recognition. International Joint Conference on Natural Language Processing(2008).

[34] S.M. Mohammad , P. Turney.Crowdsourcing a word-emotion association lexicon. Comput. Intell. 29 (3) (2013) 436–465.

[35] A.Bandhakavi, N.Wiratunga, D.Padmanabhan, S.Massie.Lexicon based feature extraction for emotion text classification.Pattern Recognition Letters 93 (2017) 133–142.

[36] M.Sap,G.Park,J.Eichstaedt,M.Kern,D.Stillwell,M.Kosinski,L.Ungar,H.Schwartz.Developing age and gender predictive lexica over social media.Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP)(2014)1146–1151.

Abstract

In recent years with the expansion of data mining and text mining, A lot of attention has focused on emotional analysis.Application of emotion detection in marketing,politics, psychology and Artificial intelligence made it more popular. Essentially, emotions in humans are characterized by various factors. Such as facial expression, blood pressure, heart rate and etc. In this project, we focus on the recognition of feelings from the text. By expanding the Internet and providing appropriate and broad public spaces To share comments and feelings, Now much text data is available to researchers. Certainly, this volume of information is impossible to study manually. So, inevitably, we have to look for methods and tools which can automatically analyze texts emotionally. In this project, We purpose tools and algorithms to solve this problem and compare them in terms of efficiency.

College of Science
School of Mathematics, Statistics, and Computer Science

Emotion Detection based on Text

Maryam Rastegari

Supervisor: Dr. Hedieh Sajedi

A thesis submitted to Graduate Studies Office
in partial fulfillment of the requirements for the degree of
B.Sc. in
Computer Science

2018