فهرست مطالب

۱ مقدمه 1
۲ مروری بر کارهای پیشین 3
۲.۱ معرفی 3
۲.۲ تجزیه و تحلیل احساسات 6
۲.۲.۱ سطوح تحلیل احساسات 7
۲.۳ تشخیص احساسات 8
۲.۳.۱ مدل های هیجانی / نظریه‌های هیجانی 8
۲.۴ مروری بر روش‌های پیشین 14
۲.۴.۱ شناختن داده ها 14
۲.۴.۲ پیش پردازش داده‌ها 16
۲.۴.۳ استخراج ویژگی 17
۲.۴.۴ تکنیک های تجزیه وتحلیل احساسات و تشخیص احساسات 19
۲.۴.۵ تکنیک های تحلیل احساسات 20
۲.۴.۶ تکنیک های تشخیص احساسات 25
۳ روش پیشنهادی 29
۳.۱ شناختن داده‌ها 30
۳.۲ پیش پردازش داده‌ها 32
۳.۲.۱ حذف کردن url 32
3.2.2 حذف mentions 34
۳.۲.۳ تمیز کردن متن 36
۳.۳ تقسیم داده‌ها 39
۳.۴ ساختن مدل 42
۳.۵.۱ SVC 44
۳.۵.۲ Random Forest 47
۳.۵.۳ XGBoost 50
۳.۵.۴ Naive Bayes 53
۴ ارزیابی روش 56
۴.۱ accuracy 56
۴.۲ Recall 56
۴.۳ F1-Score 57
۴.۴ Specificity 57
۴.۵ False Positive Rate (FPR) 57
۴.۶ False Negative Rate (FNR) 57
۴.۷ Area Under the Receiver Operating Characteristic (ROC AUC) 58
۵ نتایج 60
۶ جمع‌بندی و کارهای آینده 61
7 مراجع 62

فهرست شکلها

شکل 1: مدل بعدی احساسات.....11

شکل 2: تصویرسازی انواع مدل های عاطفی با برخی حالات روانی.....13

شکل 3: تکنیک‌هایی برای تجزیه و تحلیل احساسات و تشخیص احساسات. .20

فهرست جدولها

جدول 1: مدل های هیجانی که توسط روانشناسان مختلف تعریف شده است..

12

۱ مقدمه

در عصر دیجیتال امروزی، حجم وسیعی از داده‌های متنی موجود به صورت آنلاین، پنجره‌ای به احساسات، نظرات و احساسات جمعی افراد و جوامع می‌دهد. تجزیه و تحلیل احساسات، همچنین به عنوان نظر کاوی شناخته می شود، یک تکنیک پردازش زبان طبیعی (NLP) است که به دنبال استخراج و درک احساسات بیان شده در متن است. نقش مهمی در رمزگشایی اینکه آیا یک قطعه متن معین احساسات مثبت، منفی یا خنثی را منتقل می کند، ایفا می کند.

تجزیه و تحلیل احساسات به ابزاری قدرتمند با کاربرد در زمینه های مختلف، از تجارت و بازاریابی گرفته تا نظارت بر رسانه های اجتماعی و تحلیل سیاسی تبدیل شده است. با خودکار کردن فرآیند تشخیص احساسات، سازمان‌ها می‌توانند بینش‌های ارزشمندی به دست آورند، تصمیمات مبتنی بر داده‌ها را اتخاذ کنند و به بازخورد مشتریان پاسخ مؤثری دهند و در نهایت محصولات، خدمات و استراتژی‌های خود را بهبود بخشند.

در این عصر دیجیتال، که داده های متنی فراوان است و ارتباطات عمدتاً مبتنی بر متن است، تجزیه و تحلیل احساسات به عنوان ابزاری حیاتی برای درک چشم انداز احساسی دنیای آنلاین است. این نمای کلی مقدماتی زمینه را برای کاوش در پیچیدگی ها، روش شناسی ها و کاربردهای دنیای واقعی تحلیل احساسات فراهم می کند.

با رشد پلتفرم های مختلف در زمینه شبکه های اجتماعی این پلتفرم های به ابـزاری قدرتمند بـرای افـراد تبدیل شده اند تا افکار، نظرات و احساسات خود را در مورد مسائل مختلف بیان کننـد. ما در این پـروژه قصد داریم تکنیک های مختلف تحلیل احساسات را به کار گرفته و احساسـات را در ۱۲ گـروه مختلف طبقه بندی کنیم

این تحقیق با یک مرحله پیش پردازش کامل شروع می شود، شامل حـذف URL ها و اشـاره هـا، پـاک کردن متن، تبدیل آن به حروف کوچک، حذف فضاهای خالی اضافی، نشانه گذاری متن، حذف کلمـات توقف، علائم نقطه گذاری، و اعداد، و در نهایت واژه سازی یا ریشه کـردن کلمـات. . هـدف این مراحل پیش پردازش آماده سازی داده های متنی برای تجزیه و تحلیل بعدی است.

مرحله بعدی شامل رمزگذاری برچسب های احساسات برای تسهیل آموزش و ارزیابی مدل هـای تحلیل احساسات است. سپس داده ها به مجموعه های آموزشی و آزمایشی برای توسعه و ارزیابی مـدل تقسـیم می شوند. برای نمایش موثر داده‌های متنی، از تکنیک بردارسازی TF-IDF استفاده می‌شـود که اهمیت اصطلاحات را در اسناد نشان می‌دهد.

چندین مدل یادگیری ماشـین در این تحقیق اسـتفاده شـده اسـت، از جمله طبقه‌بنـدی بـردار پشـتیبانی (SVC)، جنگل تصادفی، XGBoost و Naive Bayes. این مدل‌ها برای طبقه‌بندی احساسات بیان‌شده در پیام‌های توییتر بر روی مجموعه داده‌ها آموزش داده شده و آزمایش می‌شوند.

ارزیابی عملکرد مدل‌ها با محاسبه دقت با استفاده از طبقه‌بنـدی_گـزارش انجـام می‌شـود، که بینش‌هـایی درباره دقت، یادآوری، امتیاز F۱ و سایر معیارهـای ارزیـابی بـرای هر کلاس احساسی ارائه می‌دهـد. این

تجزیه و تحلیل به ارزیابی اثربخشی و محدودیت‌های مدل‌ها در گرفتن تفاوت‌های ظریف احساسات در پیام‌های توییتر کمک می‌کند.

یافته های این تحقیق به حوزه تحلیل احساسات، به ویژه در زمینه پیام های شبکه اجتماعی توییتر کمک می کند. این مطالعه بینش‌هایی را در مورد عملکرد مدل‌های مختلف یادگیری ماشین و توانایی آنها در طبقه‌بندی احساسات در توییت‌هایی که با ۱۲ احساس مختلف حاشیه‌نویسی شده‌اند، ارائه می‌کند. نتایج در زمینه هایی مانند نظارت بر رسانه های اجتماعی، مدیریت شهرت برند و تحلیل افکار عمومی پیامدهای عملی دارند.

۲ مروری بر کارهای پیشین

۲.۱ معرفی

درک زبان انسانی و تولید زبان انسانی دو جنبه پردازش زبان طبیعی (NLP) هستند. با این حال، اولی به دلیل ابهامات در زبان طبیعی دشوارتر است. تشخیص گفتار، خلاصه‌سازی اسناد، پاسخ به سؤال، سنتز گفتار، ترجمه ماشینی و سایر کاربردها همگی از NLP استفاده می‌کنند (ایتانی و همکاران.۲۰۱۷) . دو حوزه مهم پردازش زبان طبیعی، تحلیل احساسات و تشخیص احساسات است. اگرچه گاهی اوقات این دو نام به جای یکدیگر استفاده می شوند، اما از چند جنبه با هم تفاوت دارند. تحلیل احساسات وسیله ای برای ارزیابی مثبت، منفی یا خنثی بودن داده ها است.

در مقابل، تشخیص احساسات ابزاری برای شناسایی انواع احساسات انسانی متمایز مانند خشمگین، شاد و یاافسرده. «تشخیص احساسات»، «محاسبات عاطفی»، «تحلیل احساسات» و «شناسایی احساسات» همگی عباراتی هستند که گاهی به جای یکدیگر استفاده می‌شوند (Munezero et al.۲۰۱۴). از زمانی که خدمات اینترنتی بهبود یافته است، مردم از رسانه های اجتماعی برای انتقال احساسات خود استفاده می کنند. در رسانه های اجتماعی، مردم آزادانه احساسات، استدلال ها، نظرات خود را در مورد طیف گسترده ای از موضوعات بیان می کنند. علاوه بر این، بسیاری از کاربران در سایت‌های مختلف تجارت الکترونیک، محصولات و خدمات مختلف را بازخورد و بررسی می‌کنند. رتبه‌بندی‌ها و بررسی‌های کاربران در چندین پلتفرم، فروشندگان و ارائه‌دهندگان خدمات را تشویق می‌کند تا سیستم‌ها،

کالاها یا خدمات فعلی خود را بهبود بخشند. امروزه تقریباً هر صنعت یا شرکتی در حال گذراندن دوره‌ای از انتقال دیجیتالی است که منجر به افزایش حجم عظیمی از داده‌های ساختاریافته و بدون ساختار می‌شود. وظیفه بزرگ شرکت‌ها تبدیل داده‌های بدون ساختار به بینش‌های معنادار است که می‌تواند به آنها در تصمیم‌گیری کمک کند (احمد و همکاران.۲۰۲۰)

به عنوان مثال، در دنیای تجارت، فروشندگان از پلتفرم های رسانه های اجتماعی مانند اینستاگرام، یوتیوب، توییتر و فیس بوک برای پخش اطلاعات در مورد محصول خود و جمع آوری کارآمد بازخورد مشتریان استفاده می کنند (Agbehadji and Ijabad-eniyi).۲۰۲۱). بازخورد فعال مردم نه تنها برای بازاریابان کسب‌وکار برای سنجش رضایت مشتری و پیگیری رقابت، بلکه برای مصرف‌کنندگانی که می‌خواهند قبل از خرید درباره یک محصول یا خدمات بیشتر بدانند ارزشمند است.آی تی. تجزیه و تحلیل احساسات به بازاریابان کمک می کند تا دیدگاه های مشتریان خود را بهتر درک کنند تا تغییرات لازم را در محصولات یا خدمات خود ایجاد کنند (Jang et al.;۲۰۱۳ الاجراوی و همکاران۲۰۲۱). در هر دو حالت پیشرفته و در حال ظهور کشورها، تأثیر احساسات تجاری و مشتریان بر عملکرد بازار سهام را می توان مشاهده کرد. علاوه بر این، ظهور رسانه‌های اجتماعی تعامل سرمایه‌گذاران در بازار سهام را آسان‌تر و سریع‌تر کرده است. در نتیجه، احساسات سرمایه‌گذاران بر تصمیم‌های سرمایه‌گذاری آنها تأثیر می‌گذارد که می‌تواند به سرعت در شبکه گسترش پیدا کند و بزرگ‌تر شود و بازار سهام تا حدی تغییر کند (احمد).۲۰۲۰). در نتیجه، تحلیل احساسات و عواطف، روشی را که ما برای کسب و کار انجام می دهیم تغییر داده است (بهاردواج و همکاران،۲۰۱۵).

در بخش مراقبت های بهداشتی، رسانه های اجتماعی آنلاین مانند توییتر به منابع ضروری اطلاعات مرتبط با سلامتی تبدیل شده اند که توسط متخصصان مراقبت های بهداشتی و شهروندان ارائه می شود. برای مثال، مردم افکار، نظرات و احساسات خود را در مورد همه‌گیری کووید-۱۹ به اشتراک گذاشته‌اند (گارسیا و برتون)۲۰۲۱). به بیماران دستور داده شد که از عزیزان خود جدا بمانند که به سلامت روان آنها آسیب می رساند. برای نجات بیماران از مسائل مربوط به سلامت روانی مانند افسردگی، پزشکان بهداشت باید از تجزیه و تحلیل خودکار احساسات و عواطف استفاده کنند (سینگ و همکاران.۲۰۲۱). مردم معمولاً احساسات یا باورهای خود را در سایت‌ها از طریق پست‌هایشان به اشتراک می‌گذارند، و

اگر فردی افسرده به نظر می‌رسد، مردم می‌توانند برای کمک به او مراجعه کنند، بنابراین از بدتر شدن شرایط سلامت روان جلوگیری می‌کنند.

تجزیه و تحلیل احساسات و عواطف نقش مهمی در بخش آموزش، هم برای معلمان و هم برای دانش آموزان ایفا می کند. کارآمدی یک معلم نه تنها با مدرک تحصیلی او تعیین می شود، بلکه با اشتیاق، استعداد و فداکاری او نیز تعیین می شود. گرفتن بازخورد به موقع از دانش آموزان موثرترین تکنیک برای معلم برای بهبود رویکردهای تدریس است (سان–گیتا و پرابها). ۲۰۲۰). مشاهده بازخورد متنی با پایان باز دشوار است و نتیجه گیری دستی نیز چالش برانگیز است. یافته های تجزیه و تحلیل احساسات و تحلیل احساسات به معلمان و سازمان ها در انجام اقدامات اصلاحی کمک می کند. از زمان تأسیس سایت اجتماعی، مؤسسات آموزشی به طور فزاینده ای به رسانه های اجتماعی مانند فیس بوک و توییتر برای اهداف بازاریابی و تبلیغات متکی هستند. دانشجویان و سرپرستان تحقیقات آنلاین قابل توجهی را انجام می دهند و در مورد مؤسسه، دوره ها و اساتید بالقوه اطلاعات بیشتری کسب می کنند. آنها از وبلاگ ها و دیگر انجمن های گفتگو برای تعامل با دانشجویانی که علایق مشترک دارند و ارزیابی کیفیت کالج ها و دانشگاه های احتمالی استفاده می کنند. بنابراین، استفاده از تحلیل احساسات و عواطف می تواند به دانش آموز کمک کند تا بهترین مؤسسه یا معلم را در فرآیند ثبت نام خود انتخاب کند (Archana Rao و Baglodi). ۲۰۱۷).

تحلیل احساسات و عواطف کاربردهای گسترده ای دارد و با استفاده از روش شناسی های مختلف قابل انجام است. سه نوع تکنیک تحلیل احساسات و عواطف وجود دارد: مبتنی بر واژگان، مبتنی بر یادگیری ماشینی و مبتنی بر یادگیری عمیق. هر کدام مزایا و معایب خاص خود را دارند. علیرغم تشخیص احساسات و عواطف متفاوتمحققان با چالش‌های مهمی از جمله برخورد با زمینه، تمسخر، جملاتی که چندین احساسات را منتقل می‌کنند، گسترش عامیانه وب و ابهامات واژگانی و نحوی مواجه هستند. علاوه بر این، از آنجایی که هیچ قانون استانداردی برای برقراری ارتباط احساسات در چندین پلتفرم وجود ندارد، برخی آن‌ها را با تأثیری باورنکردنی بیان می‌کنند، برخی احساسات خود را خفه می‌کنند و برخی پیام خود را به صورت منطقی ساختار می‌دهند. بنابراین، توسعه تکنیکی که بتواند به طور موثر در همه حوزه ها کار کند، چالش بزرگی برای محققان است.

در این مقاله مروری، بخش.۲، تجزیه و تحلیل احساسات و سطوح مختلف آن، تشخیص احساسات و مدل های روانی–منطقی را معرفی می کند. بخش۳مراحل متعددی را که در تجزیه و تحلیل احساسات و عواطف دخیل هستند، از جمله مجموعه داده ها، پیش پردازش متن، تکنیک های استخراج ویژگی، و رویکردهای مختلف تحلیل احساسات و عواطف را مورد بحث قرار می دهد. بخش۴چالش‌های متعددی را که محققان در طول تجزیه و تحلیل احساسات و عواطف با آن مواجه هستند، بررسی می‌کند. در نهایت، بخش.۵کار را به پایان می رساند.

۲.۲ تجزیه و تحلیل احساسات

در حال حاضر بسیاری از مردم در سراسر جهان از وبلاگ ها، انجمن ها و سایت های رسانه های اجتماعی مانند توییتر و فیس بوک برای به اشتراک گذاشتن نظرات خود با سایر نقاط جهان استفاده می کنند. رسانه های اجتماعی به یکی از موثرترین رسانه های ارتباطی موجود تبدیل شده اند. در نتیجه، مقدار زیادی داده تولید می شود که داده های بزرگ نامیده می شود و تجزیه و تحلیل احساسات برای تجزیه و تحلیل موثر و کارآمد این کلان داده معرفی شد (ناگامانجولا و پتالاکشمی ۲۰۲۰). درک احساسات کاربر برای صنعت یا سازمان بسیار مهم است. تجزیه و تحلیل احساسات، که اغلب به عنوان عقیده کاوی شناخته می شود، روشی برای تشخیص مثبت یا منفی بودن دیدگاه نویسنده یا کاربر در مورد یک موضوع است. تحلیل احساسات به عنوان فرآیند به دست آوردن اطلاعات و معناشناسی معنادار از متن با استفاده از تکنیک های پردازش طبیعی و تعیین نگرش نویسنده که ممکن است مثبت، منفی یا خنثی باشد، تعریف می شود (Onyenwe et al.۲۰۲۰). از آنجایی که هدف از تجزیه و تحلیل احساسات، تعیین قطبیت و طبقه بندی متون دارای نظر مثبت یا منفی است، محدوده کلاس مجموعه داده درگیر در تحلیل احساسات فقط به مثبت یا منفی محدود نمی شود. می تواند موافق یا مخالف باشد، خوب یا بد. همچنین می توان آن را در مقیاس ۵ درجه ای تعیین کرد: کاملاً مخالفم، مخالفم، خنثی، موافقم، یا کاملاً موافقم (پرابوو و تلوال۲۰۰۹). به عنوان مثال، (۲۰۰۹) Ye et al. تجزیه و تحلیل احساسات را روی نظرات در مقاصد اروپایی و ایالات متحده با برچسب در مقیاس ۱ تا ۵ اعمال کردند. آنها نظرات ۱ یا ۲ ستاره را با قطبیت منفی و نظرات بیش از ۲ ستاره را با قطبیت مثبت مرتبط کردند. گرابنر و همکاران (۲۰۱۲) یک واژگان مخصوص دامنه را ساخت که از نشانه هایی با ارزش احساسی

آنها تشکیل شده است. این توکن ها بودناز نظـرات مشـتریان در حـوزه گردشـگری بـرای طبقـه بنـدی احساسات به رتبه بندی های ۵ ستاره از وحشتناک تا عالی در حوزه گردشگری جمع آوری شده اسـت. علاوه بر این، تحلیل احساسات از متن می تواند در سـه سـطح مـورد بحث در بخش زیـر انجـام شـود. سالینکا (۲۰۱۵) الگوریتم های یـادگیری ماشـینی را روی مجموعـه داده Yelp اعمـال کـرد کـه شـامل بررسی ارائه دهندگان خدمات از ۱ تا ۵ است. تجزیه و تحلیل احساسات را می توان در سه سطح طبقـه بندی کرد که در بخش بعدی اشاره شد.

۲.۲.۱ سطوح تحلیل احساسات

تحلیل احساسات در سه سطح امکان پذیر است: سـطح جملـه، سـطح سـند و سـطح جنبـه. در تحلیـل احساسات در سطح جمله یا عبارت، اسناد یا پاراگراف ها بـه جملات تقسـیم می شـوند و قطـبیت هـر جملـه مشـخص می شـود (Meena و Prabhakar، ۲۰۰۷; آرولموروگـان و همکـاران، ۲۰۱۹; شیرست و همکاران، ۲۰۱۹). در سطح سند، احساس از کل سند یا رکـورد شناسـایی می شـود (Pu et al.، ۲۰۱۹). ضرورت تجزیه و تحلیل احساسات در سـطح سـند، اسـتخراج احساسـات جهـانی از متـون طولانی است که حاوی الگوهای محلی اضافی و نویز زیادی هستند. چالش برانگیزترین جنبه طبقه بندی احساسات در سطح سند، در نظر گرفتن پیوند بین کلمات و عبارات و زمینه کامل اطلاعات معنایی بـرای انعکاس ترکیب سند است (رائـو و همکـاران، ۲۰۱۸; لیـو و همکـاران ۲۰۲۰a). این امـر مسـتلزم درک عمیق‌تر ساختار درونی درونی احساسات و کلمات وابسـته اسـت (لیـو و همکـاران، ۲۰۲۰b). در سـطح جنبه، تحلیل احساسات، نظر در مورد یک جنبه یا ویژگی خاص تعیین می شود. به عنوان مثـال، سـرعت پردازنده بالا است، اما این محصول قیمت بـالایی دارد. در اینجـا سـرعت و هزینـه دو جنبـه یـا دیـدگاه هستند. سرعت در جمله اشاره شده است، از این رو جنبه صریح نامیده می شود، در حالی که هزینه یک جنبه ضمنی است. تجزیه و تحلیل احساسات سطح جنبه کمی سخت تر از دو مـورد دیگـر اسـت زیـرا شناسایی ویژگی های ضمنی دشوار است. دوی سـری نانـدینی و پـرادیپ (۲۰۲۰) الگـوریتمی را بـرای استخراج جنبه های ضمنی از اسناد بر اساس فراوانی همزمانی جنبه با شاخص ویژگی و با بهره بـرداری از رابطه بین کلمات نظری و جنبه های صریح پیشنهاد کرد. ما و همکاران (۲۰۱۹) به دو موضوع مربـوط به تحلیل سطح جنبه رسیدگی کرد: جنبه های مختلف در یک جمله با قطبیت هـای متفـاوت و مـوقعیت صریح بافت در یک جمله نظری. نویسندگان یک مدل دو مرحله ای بر اساس LSTM با مکانیزم توجه

برای حل این مسائل ایجاد کردند. آنها این مدل را بر اساس این فرض پیشنهاد کردند که کلمات بافت نزدیک به جنبه مرتبط تر هستند و نیاز به توجه بیشتری نسبت به کلمات بافت دورتر دارند. در مرحله یک، مدل از جنبه های متعدد در یک جمله یک به یک با مکانیسم توجه موقعیت بهره برداری می کند. سپس در حالت دوم، جفت‌ها (جنبه، جمله) را با توجه به موقعیت جنبه و زمینه اطراف آن شناسایی می‌کند و قطبیت هر تیم را به طور همزمان محاسبه می‌کند.همانطور که قبلاً گفته شد، تحلیل احساسات و تحلیل احساسات اغلب به جای یکدیگر توسط محققان مورد استفاده قرار می گیرند. با این حال، آنها از چند جهت متفاوت هستند. در تجزیه و تحلیل احساسات، قطبیت نگرانی اصلی است، در حالی که، در تشخیص احساسات، حالت یا خلق عاطفی یا روانی تشخیص داده می شود. تجزیه و تحلیل احساسات به طور استثنایی ذهنی است، در حالی که تشخیص احساسات عینی تر و دقیق تر است. بخش۲.۲همه چیز در مورد تشخیص احساسات را با جزئیات شرح می دهد.

۲.۳ تشخیص احساسات

احساسات جزء جدایی ناپذیر زندگی انسان هستند. این احساسات بر تصمیم‌گیری انسان تأثیر می‌گذارد و به ما کمک می‌کند تا به روشی بهتر با جهان ارتباط برقرار کنیم. تشخیص عواطف، که به عنوان تشخیص احساسات نیز شناخته می شود، فرآیند شناسایی احساسات یا عواطف مختلف یک فرد (به عنوان مثال، شادی، غم یا خشم) است. محققان در چند سال گذشته سخت کار کرده اند تا تشخیص احساسات را خودکار کنند. با این حال، برخی از فعالیت‌های فیزیکی مانند ضربان قلب، لرز دست‌ها، تعریق و زیر و بمی صدا نیز وضعیت عاطفی فرد را منتقل می‌کنند (Kratzwald et al.۲۰۱۸)، اما تشخیص احساسات از روی متن بسیار سخت است. علاوه بر این، هر روز که می گذرد، ابهامات مختلف و اصطلاحات یا اصطلاحات عامیانه جدید معرفی می شود، تشخیص احساسات از متن را چالش برانگیزتر می کند. علاوه بر این، تشخیص احساسات فقط به شناسایی شرایط روانی اولیه (شادی، غمگین، خشم) محدود نمی شود. در عوض، بسته به مدل احساسی، به مقیاس ۶ یا ۸ می رسد.

۲.۳.۱ مدل های هیجانی / نظریه‌های هیجانی

در زبان انگلیسی، کلمه "احساس" در قرن هفدهم به وجود آمد که از کلمه فرانسوی "emo-tion" به معنای اختلال جسمانی گرفته شده است. قبل از قرن نوزدهم، اشتیاق، اشتها و محبت ها به عنوان حالت های روانی طبقه بندی می شدند. در قرن نوزدهم، واژه «هیجان» یک اصطلاح روانشناختی در نظر گرفته شد (دیکسون۲۰۱۲). در روانشناسی، حالات پیچیده احساس منجر به تغییر در افکار، اعمال، رفتار و شخصیت می شود که به آن احساسات می گویند. به طور کلی، مدل‌های روان‌شناختی یا هیجانی به دو دسته طبقه‌بندی می‌شوند: بعدی و مقوله‌ای.

*مدل عاطفه بعدی*این مدل احساسات را بر اساس سه پارامتر نشان می‌دهند ظرفیت، برانگیختگی و قدرت (باکر و همکاران.۲۰۱۴. والانس به معنای قطبیت است و برانگیختگی به معنای هیجان انگیز بودن یک احساس است. به عنوان مثال، خوشحال بودن بیشتر هیجان انگیز است تا خوشحال. قدرت یا تسلط به معنای محدودیت بر احساسات است. این پارامترها موقعیت حالات روانی را در فضای دو بعدی تعیین می کنند، همانطور که در شکل ۱ نشان داده شده است.۱.

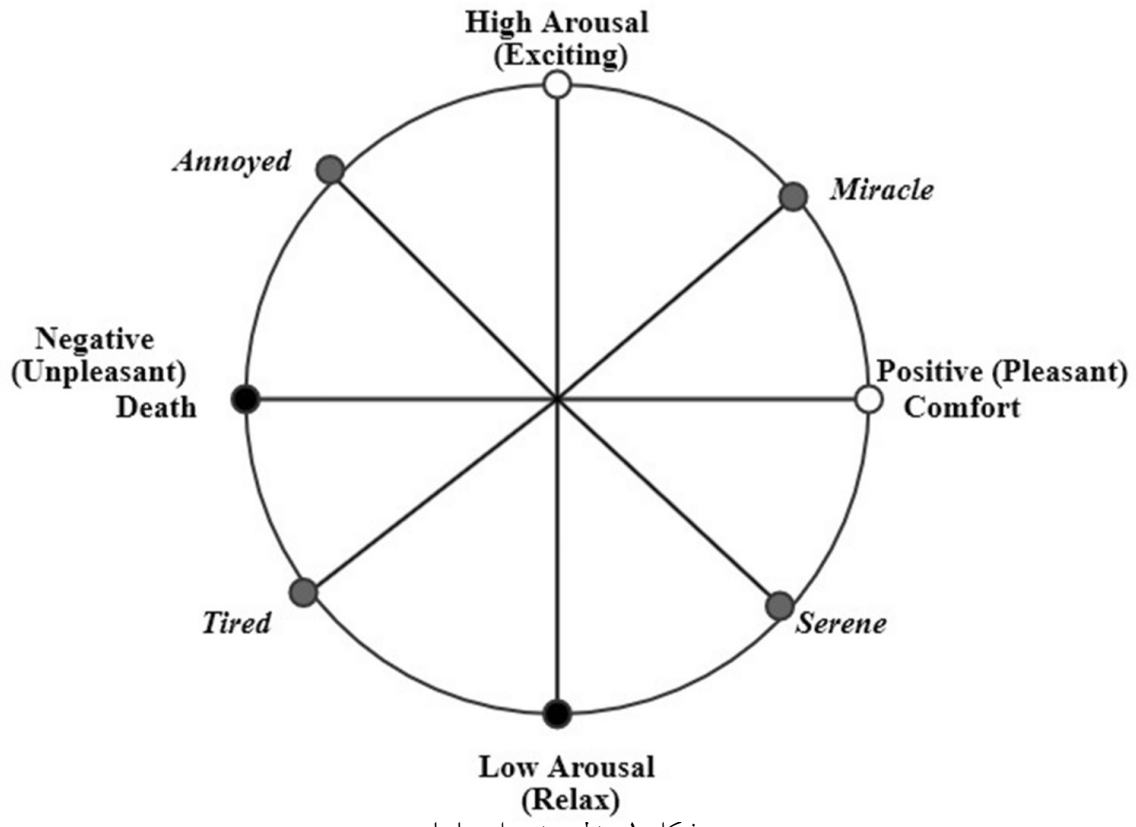
مدل احساسات طبقه بندی شده

در مدل مقوله ای، احساسات به صورت گسسته تعریف می شوند، مانند خشم، شادی، غم و ترس. بسته بهدر مدل طبقه‌بندی خاص، احساسات به چهار، شش یا هشت دسته طبقه‌بندی می‌شوند.

جدول۱مدل‌های هیجانی متعددی را نشان می‌دهد که ابعادی و مقوله‌ای هستند. در حوزه تشخیص احساسات، اکثر محققان مدل احساسات اکمن و پلاچیک را اتخاذ کردند. حالت های احساسی تعریف شده توسط مدل ها مجموعه ای از برچسب ها را تشکیل می دهد که برای حاشیه نویسی جملات یا اسناد استفاده می شود. Batbaatar و همکاران. (۲۰۱۹بکر و همکاران. (۲۰۱۷جین و همکاران. (۲۰۱۷) شش احساس اساسی اکمن را پذیرفت. سای لوناز و الحاج (۲۰۱۹) از مدل های اکمن برای حاشیه نویسی توییت ها استفاده کرد. برخی از محققان با گسترش مدل با یک یا دو حالت اضافی، از مدل های احساسی سفارشی استفاده کردند. رابرتز و همکاران (۲۰۱۲) از مدل اکمن برای حاشیه نویسی توییت ها با حالت "عشق" استفاده کرد. احمد و همکاران (۲۰۲۰) چرخ احساسات مدل‌سازی شده

توسط پلوچیک را برای برچسب‌گذاری جملات هندی با ۹ حالت مختلف مدل پلوچیک، کاهش سردرگمی معنایی، در میان کلمات دیگر، اتخاذ کرد. حالت های مدل پلاچیک و اکمن نیز در واژگان دست ساز مختلف مانند WordNet-Affect استفاده می شود (Strapparava et al.۲۰۰۴) و NRC (محمد و تورنی۲۰۱۳) واژگان کلمه–احساس. لوبرت و پارلامیس (۲۰۱۹) به مدل Shaver به دلیل ساختار سلسله مراتبی سه سطحی احساسات اشاره کرد. ظرفیت یا قطبیت در سطح اول ارائه می شود، سپس سطح دوم شامل پنج احساس است و سطح سوم ۲۴ حالت احساسی گسسته را نشان می دهد. برخی از محققان به هیچ مدلی اشاره نکردند و مجموعه داده را به سه احساس اصلی طبقه بندی کردند: خوشحال، غمگین یا عصبانی.

شکل۲حالت های عاطفی متعددی را که در مدل های مختلف یافت می شود به تصویر می کشد. این حالت ها با در نظر گرفتن مدل پلاچیک به عنوان مدل پایه بر روی چهار محور ترسیم می شوند. رارایج ترین حالت های هیجانی مورد استفاده در مدل های مختلف شامل خشم، ترس، شادی، تعجب و انزجار است که در شکل بالا نشان داده شده است. از شکل می توان دریافت که احساسات در دو طرف محور همیشه مخالف یکدیگر نخواهند بود. مثلاً غم و شادی متضاد یکدیگرند، اما خشم مخالف ترس نیست.

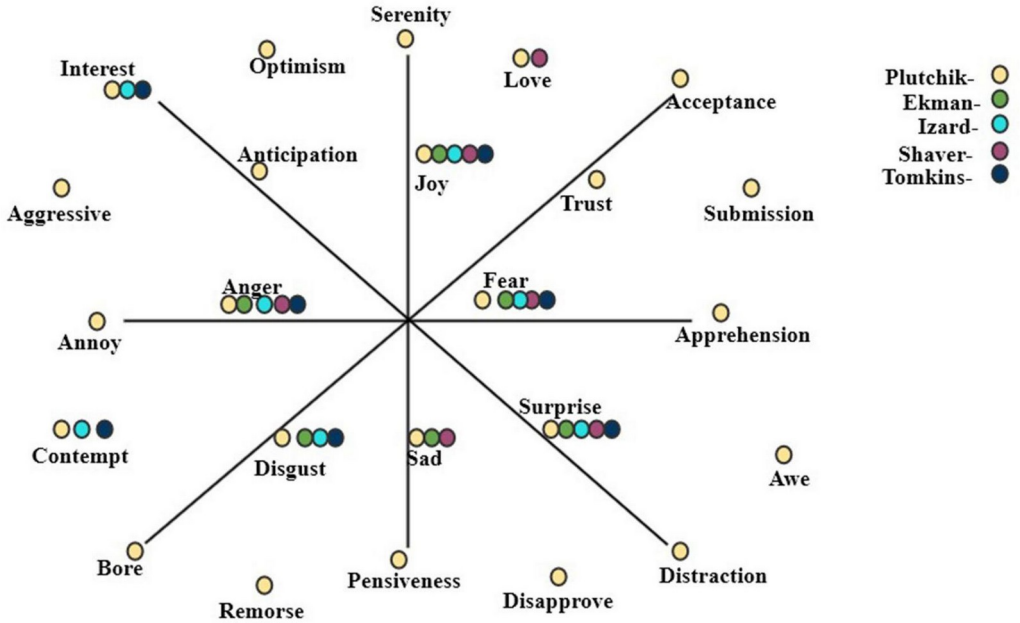


شکل ۱: مدل بعدی احساسات

جدول ۱: مدل های هیجانی که توسط روانشناسان مختلف تعریف شده است

Emotions	Type of model	No, of states	Psychological states	Representations	Discussion
Ekman model (Ekman 1992)	Categorical	6	Anger, disgust, fear, joy, sadness, surprise	–	Ekman's model consisted of six emotions, which act as a base for other emotion models like Plutchik model
Plutchik Wheel of Emotions (Plutchik 1982)	Dimensional	-	Joy, pensiveness, ecstasy, acceptance, sadness, fear, interest, rage, admiration, amazement, anger, vigilance boredom, annoyance, submission, serenity, apprehension, contempt, surprise, disapproval, distraction, grief, loathing, love, optimism, aggressiveness, remorse, anticipation, awe, terror, trust, disgust	wheel	Plutchik considered two types of emotions: basic (Ekman model + Trust +Anticipation) and mixed emotions (made from the combination of basic emotions). Plutchik represented emotions on a colored wheel
Izard model (Izard 1992)	–	10	Anger, contempt, disgust, anxiety, fear, guilt, interest, joy, shame, surprise	–	–
Shaver model (Shaver et al. 1987)	Categorical	6	Sadness, joy, anger, fear, love, surprise	tree	Shaver represented the primary, secondary and tertiary emotions in a hierarchal manner. The top-level of the tree presents these six emotions
Russell's circumplex model (Russell 1980)	Dimensional	-	Sad, satisfied, Afraid, alarmed, frustrated, angry, happy, gloomy, annoyed, tired, relaxed, glad, aroused, astonished, at ease, tense, miserable, content, bored, calm, delighted, excited, depressed, distressed, serene, droopy, pleased, sleepy	–	Emotions are presented over the circumplex model
Tomkins model (Tomkins and		9		–	

McCarter 1964)	Categorical		Disgust, surprise-Startle, anger-rage, anxiety, fear-terror, contempt, joy, shame, interest-Excitement		Tomkins identified nine different emotions out of which six emotions are negative. Most of the emotions are defined as a pair
Lövheim Model (Lövheim 2012)	Dimensional	-	Anger, contempt, distress, enjoyment, terror, excitement, humiliation, startle	cube	Lövheim arranged the emotions according to the amount of three substances (Noradrenaline, dopamine and Serotonin) on a ۳-D cube



شکل ۲: تصویرسازی انواع مدل های عاطفی با برخی حالات روانی

۲.۴ مروری بر روش‌های پیشین

در پردازش احساسات به صورت کلی مراحل زیر طی می‌شوند.

برای پیاده‌سازی هرکدام از این مراحل روش‌های بسیار متنوعی وجود دارند که در ادامه شرح داده میشوند.

۲.۴.۱ شناختن داده ها

قبل از انجام هر مرحله پیش پردازش در پردازش زبان طبیعی (NLP) درک ویژگی ها و ویژگی های مجموعه داده از اهمیت بالایی برخوردار است. مجموعه داده، که به عنوان پایه ای برای تجزیه و تحلیل عمل می کند، بینش های ارزشمندی را ارائه می دهد که تصمیمات پیش پردازش را هدایت می کند و در نهایت بر کیفیت نتایج تأثیر می گذارد.

قبل از هر چیز، دانستن مجموعه داده ها به محققان اجازه می دهد تا درک جامعی از داده های متنی که با آنها کار می کنند به دست آورند. با بررسی مجموعه داده ها، می توان زبان مورد استفاده، دامنه یا موضوع متون، وجود اصطلاحات خاص یا اصطلاحات عامیانه و ساختار و قالب کلی داده ها را شناسایی کرد. این دانش به تطبیق مراحل پیش پردازش با زمینه خاص کمک می‌کند و اطمینان حاصل می‌کند که تحلیل‌ها و مدل‌های بعدی به خوبی با مجموعه داده سازگار هستند.

علاوه بر این، درک مجموعه داده ها به شناسایی چالش ها و مسائل بالقوه ای که ممکن است در طول پیش پردازش ایجاد شود کمک می کند. به عنوان مثال، وجود داده‌های پر سر و صدا، مانند اشتباهات تایپی، جملات ناقص، یا ناهماهنگی در قالب‌بندی، ممکن است به تکنیک‌های ویژه‌ای نیاز داشته باشد. به همین ترتیب، مجموعه داده ممکن است شامل کاراکترهای خاص، ایموجی ها یا نمادهای غیر

استاندارد باشد که باید به درستی مورد توجه قرار گیرند. با آگاهی از این چالش ها از قبل، محققان می توانند خطوط لوله پیش پردازش قوی ایجاد کنند که این مسائل را به طور موثر کاهش داده یا به آنها رسیدگی کند.

علاوه بر این، درک مجموعه داده ها امکان تصمیم گیری آگاهانه در مورد انتخاب و کاربرد تکنیک های پیش پردازش را فراهم می کند. تکنیک‌های مختلف، مانند نشانه‌سازی، حذف کلمه توقف، ریشه‌یابی، واژه‌سازی یا عادی‌سازی، بسته به ویژگی‌های آن می‌توانند تأثیرات متفاوتی بر داده‌ها داشته باشند. به عنوان مثال، اگر مجموعه داده شامل محتوای تولید شده توسط کاربر از پلتفرم‌های رسانه‌های اجتماعی باشد، ممکن است به مراحل پیش‌پردازش خاصی نیاز داشته باشد تا هشتگ‌ها، اشاره ها یا URLها را به درستی مدیریت کند. با دانستن مجموعه داده، محققان می توانند تصمیمات آگاهانه ای در مورد اینکه کدام تکنیک ها را به کار گیرند و چگونه آنها را برای مجموعه داده خاص بهینه کنند، اتخاذ کنند که منجر به نتایج دقیق تر و معنی داری شود.

علاوه بر این، درک مجموعه داده، محققان را قادر می‌سازد تا سوگیری‌ها یا محدودیت‌های بالقوه‌ای را که ممکن است بر تحلیل تأثیر بگذارد، شناسایی کنند. سوگیری‌ها می‌توانند از فرآیند جمع‌آوری مجموعه داده‌ها ناشی شوند، مانند سوگیری‌های نمونه‌گیری یا سوگیری‌های جمعیتی در محتوای تولید شده توسط کاربر. با آگاهی از این سوگیری ها، محققان می توانند اقدامات مناسبی را برای رسیدگی یا کاهش آنها در طول پیش پردازش و تجزیه و تحلیل انجام دهند و از یک نتیجه عینی تر و نماینده تر اطمینان حاصل کنند.

در نتیجه، دانستن مجموعه داده قبل از پیش پردازش برای اطمینان از دقت، ارتباط و قابلیت اطمینان تجزیه و تحلیل بعدی حیاتی است. با به دست آوردن درک کامل از مجموعه داده، محققان می توانند

مراحل پیش پردازش را با زمینه خاص تنظیم کنند، به چالش ها و سوگیری ها بپردازند، و تصمیمات آگاهانه ای در مورد تکنیک های مورد استفاده اتخاذ کنند. در نهایت، این دانش کیفیت پیش پردازش را افزایش می‌دهد و پایه محکمی برای تحلیل و مدل‌سازی موفق NLP ایجاد می‌کند.

۲.۴.۲ پیش پردازش داده‌ها

در رسانه‌های اجتماعی، مردم معمولاً احساسات و عواطف خود را به روش‌هایی بدون زحمت بیان می‌کنند. در نتیجه، داده‌های به‌دست‌آمده از پست‌ها، ممیزی‌ها، نظرات، اظهارات و انتقادات این پلت‌فرم رسانه‌های اجتماعی بسیار ساختارمند نیستند و تجزیه و تحلیل احساسات و احساسات را برای ماشین‌ها دشوار می‌کنند. در نتیجه، پیش پردازش یک مرحله حیاتی در پاکسازی داده ها است، زیرا کیفیت داده ها به طور قابل توجهی روی بسیاری از رویکردهای پیش پردازش تأثیر می گذارد.

سازماندهی یک مجموعه داده نیاز به پیش پردازش دارد، از جمله توکن سازی، حذف کلمه توقف، برچسب گذاری POS و غیره (عبدی و همکاران.۲۰۱۹; باسکار و همکاران۲۰۱۵). برخی از این تکنیک‌های پیش پردازش می‌توانند منجر به از دست رفتن اطلاعات حیاتی برای تجزیه و تحلیل احساسات و عواطف شوند که باید مورد توجه قرار گیرند.

توکن‌سازی فرآیند تجزیه کل سند یا پاراگراف یا فقط یک جمله به تکه‌هایی از کلمات به نام نشانه است (ناگاراجان و گاندی).۲۰۱۹). به عنوان مثال، جمله "این مکان بسیار زیبا است" را در نظر بگیرید و پس از توکنیزاسیون، تبدیل به "این"، "مکان"، "بسیار زیبا" است. عادی سازی متن برای دستیابی به یکنواختی در داده ها با تبدیل متن به فرم استاندارد، تصحیح املای کلمات و غیره ضروری است (آهوجا و همکاران.۲۰۱۹).

کلمات غیرضروری مانند مقاله ها و برخی حرف های اضافه که به شناخت احساسات و تحلیل احساسات کمک نمی کنند باید حذف شوند. به عنوان مثال، کلمات توقفی مانند "an"، "at"، "is"، "the" هیچ ربطی به احساسات ندارند، بنابراین برای جلوگیری از محاسبات غیرضروری، باید حذف شوند (Bhaskar et al.۲۰۱۵; عبدی و همکاران۲۰۱۹). برچسب گذاری POS راهی برای شناسایی بخش های مختلف گفتار در یک جمله است. این مرحله برای یافتن جنبه‌های مختلف یک

جمله مفید است که عموماً با اسم‌ها یا عبارات اسمی توصیف می‌شوند در حالی که احساسات و عواطف با صفت‌ها منتقل می‌شوند (سان و همکاران.۲۰۱۷).

ریشه زایی و ریشه یابی دو مرحله مهم پیش پردازش هستند. در stemming، کلمات با کوتاه کردن پسوندها به شکل ریشه خود تبدیل می شوند. به عنوان مثال، اصطلاحات «استدلال» و «استدلال» به «استدلال» تبدیل می شود. این فرآیند محاسبه ناخواسته جملات را کاهش می دهد (کراتزوالد و همکاران ۲۰۱۸; آکیلاندسواری و جوتی۲۰۱۸).ـ Lemmatization شامل تجزیه و تحلیل مورفولوژیکی برای حذف پایان های عطفی از یک نشانه و تبدیل آن به کلمه پایه لم است (قنبری–ادیوی و مصلح). ۲۰۱۹). به عنوان مثال، اصطلاح "گرفتار" به "گرفتن" تبدیل می شود (آهوجا و همکاران.۲۰۱۹). Symeonidis و همکاران. (۲۰۱۸) عملکرد چهار مدل یادگیری ماشین را با مطالعه ترکیبی و فرسایشی تکنیک های مختلف پیش پردازش روی دو مجموعه داده، یعنی SS-Tweet و SemEval مورد بررسی قرار داد. نویسندگان به این نتیجه رسیدند که حذف اعداد و واژه‌سازی دقت را افزایش می‌دهد، در حالی که حذف علائم نگارشی بر دقت تأثیری ندارد.

۲.۴.۳ استخراج ویژگی

دستگاه متن را بر حسب اعداد درک می کند. فرآیند تبدیل یا نگاشت متن یا کلمات به بردارهای با ارزش واقعی را واژه برداری یا جاسازی کلمه می نامند. این یک تکنیک استخراج ویژگی است که در آن یک سند به جملاتی تقسیم می شود که بیشتر به کلمات تقسیم می شوند. پس از آن، نقشه یا ماتریس ویژگی ساخته می شود. در ماتریس به دست آمده، هر ردیف نشان دهنده یک جمله یا سند است در حالی که هر ستون ویژگی یک کلمه را درفرهنگ لغت، و مقادیر موجود در سلول های نقشه ویژگی به طور کلی تعداد کلمه در جمله یا سند را نشان می دهد. برای انجام استخراج ویژگی، یکی از ساده‌ترین روش‌های مورد استفاده، "Bag of Words" (BOW) است که در آن یک بردار با طول ثابت از شمارش تعریف می‌شود که در آن هر ورودی با یک کلمه در فرهنگ لغت از پیش تعریف‌شده کلمات مطابقت دارد. . اگر کلمه در یک جمله در فرهنگ لغت از پیش تعریف شده وجود نداشته باشد، تعداد ۰ اختصاص داده می شود، در غیر این صورت بسته به تعداد دفعات ظاهر شدن آن در جمله، تعداد آن بیشتر یا مساوی ۱ است. به همین دلیل است که طول بردار همیشه با کلمات موجود در فرهنگ لغت برابر است. مزیت این تکنیک اجرای آسان آن است اما دارای اشکالات قابل توجهی است زیرا منجر به

یک ماتریس پراکنده می شود، ترتیب کلمات را در جمله از دست می دهد.۲۰۱۷; عبدی و همکاران ۲۰۱۹). به عنوان مثال، برای نشان دادن متن "آیا از خواندن لذت می برید" از فرهنگ لغت از پیش تعریف شده I, Hope, you, are, enjoying, reading خواهد بود (۰,۰,۱,۱,۱,۱). با این حال، این نمایش ها را می توان با پیش پردازش متن و با استفاده از TF-IDF، n-gram بهبود بخشید.

روش N-gram یک گزینه عالی برای حل ترتیب کلمات در نمایش برداری جمله است. در یک نمایش برداری n-gram، متن به عنوان یک همکاری از n-gram منحصر به فرد به معنای گروه هایی از n واژه یا کلمه مجاور نشان داده می شود. مقدار n می تواند هر عدد طبیعی باشد. به عنوان مثال، جمله "آموزش برای همیشه لمس کردن است" را در نظر بگیرید و ۳ = n به نام تریگرام "آموزش دادن است"، "تدریس کردن است،" "لمس کردن است"، "لمس کردن است" "لمس کردن است" ایجاد می کند. یک زندگی، "یک زندگی برای همیشه". به این ترتیب می توان نظم جمله را حفظ کرد (آهوجا و همکاران۲۰۱۹) . ویژگی‌های N-gram بهتر از رویکرد BOW عمل می‌کنند، زیرا الگوهای نحوی، از جمله اطلاعات مهم (Chaffar و Inkpen) را پوشش می‌دهند.۲۰۱۱). با این حال، اگرچه n-gram ترتیب کلمات را حفظ می کند، اما ابعاد و پراکندگی داده بالایی دارد (Le and Mikolov۲۰۱۴).

اصطلاح فرکانس معکوس سند فرکانس، که معمولاً به اختصار TFIDF نامیده می شود، روش دیگری است که معمولاً برای استخراج ویژگی استفاده می شود. این روش متن را به صورت ماتریسی نشان می دهد، که در آن هر عدد مقدار اطلاعاتی را که این اصطلاحات در یک سند معین حمل می کنند را کمیت می دهد. بر این فرض ساخته شده است که اصطلاحات نادر دارای اطلاعات زیادی در سند متنی هستند (Liu et al.۲۰۱۹). فراوانی عبارت تعداد دفعاتی است که یک کلمه w در یک سند تقسیم می شود بر تعداد کل کلمات W در سند و IDF عبارت است از log (تعداد کل اسناد (N) تقسیم بر تعداد کل اسنادی که کلمه w در آنها ظاهر می شود. (ن)) (سونگبو و جین۲۰۰۸). آهوجا و همکاران (۲۰۱۹) شش تکنیک پیش پردازش را اجرا کرد و دو تکنیک استخراج ویژگی را برای شناسایی بهترین رویکرد مقایسه کرد. آنها از شش الگوریتم یادگیری ماشین استفاده کردند و از n-gram با ۲ = n و TF-

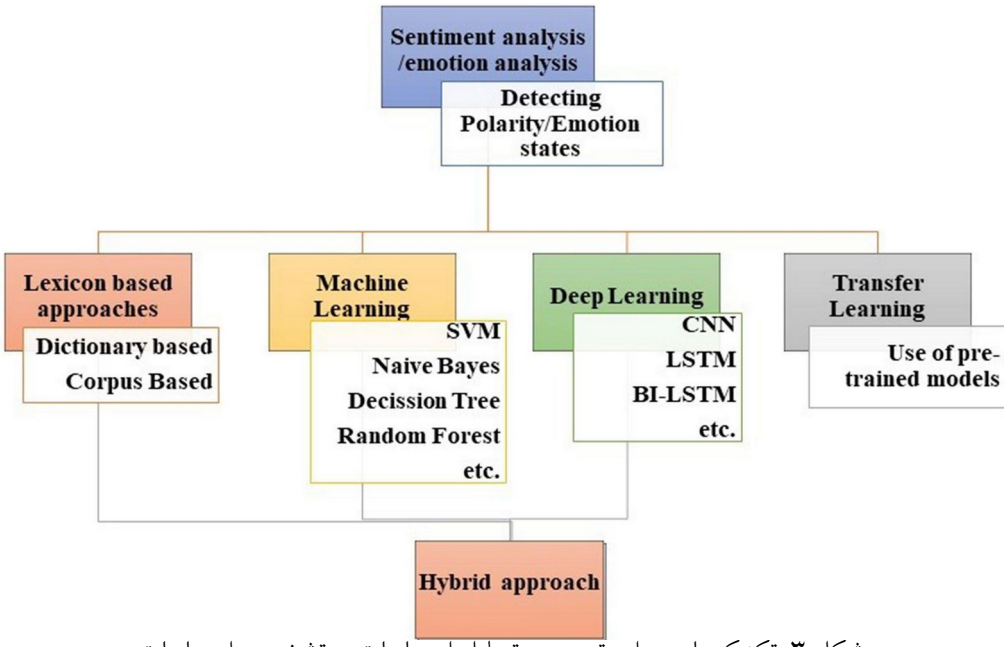
IDF برای استخراج ویژگی روی مجموعه داده SS-tweet استفاده کردند و به این نتیجه رسیدند که TF-IDF عملکرد بهتری نسبت به n-gram دارد.

در دسترس بودن حجم وسیعی از داده ها به یک شبکه یادگیری عمیق اجازه می دهد تا بازنمایی های برداری خوبی را کشف کند. استخراج ویژگی با جاسازی کلمه بر اساس شبکه های عصبی آموزنده تر است. در جاسازی کلمه مبتنی بر شبکه عصبی، کلمات با معنایی یکسان یا مرتبط با یکدیگر با بردارهای مشابه نمایش داده می شوند. این در پیش‌بینی کلمات محبوب‌تر است زیرا معنای کلمات را حفظ می‌کند. تیم تحقیقاتی گوگل به سرپرستی توماس میکولوف، مدلی به نام Word۲Vec را برای جاسازی کلمه توسعه دادند. با Word۲Vec، می‌توان برای ماشینی فهمید که نمایش برداری «ملکه» + «مونث» + «مرد» مانند بازنمایی بردار «شاه» است (سوما و همکاران.۲۰۱۹).

نمونه‌های دیگر از مدل‌های جاسازی کلمه مبتنی بر یادگیری عمیق عبارتند از GloVe که توسط محققان دانشگاه استنفورد توسعه یافته و FastText که توسط Face-book معرفی شده‌اند. آموزش بردارهای GloVe سریعتر از Word۲vec است. بردارهای FastText در مقایسه با بردارهای Word۲Vec با چندین معیار متفاوت دقت بهتری دارند. یانگ و همکاران (۲۰۱۸) ثابت کرد که انتخاب جاسازی کلمه مناسب بر اساس شبکه های عصبی می تواند منجر به پیشرفت های قابل توجهی حتی در مورد کلمات خارج از واژگان (OOV) شود. نویسندگان انواع جاسازی کلمات را که با استفاده از توییتر و ویکی‌پدیا به‌عنوان مجموعه آموزش دیده بودند، با جاسازی کلمه TF-IDF مقایسه کردند.

۲.۴.۴ تکنیک های تجزیه وتحلیل احساسات و تشخیص احساسات

شکل۴تکنیک های مختلفی را برای تجزیه و تحلیل احساسات و تشخیص احساسات ارائه می دهد که به طور گسترده به رویکرد مبتنی بر فرهنگ لغت، رویکرد مبتنی بر یادگیری ماشین، رویکرد مبتنی بر یادگیری عمیق طبقه بندی می شود. رویکرد ترکیبی ترکیبی از رویکردهای آماری و یادگیری ماشینی برای غلبه بر اشکالات هر دو رویکرد است. یادگیری انتقالی نیز زیرمجموعه ای از یادگیری ماشین است که امکان استفاده از مدل از پیش آموزش دیده را در سایر حوزه های مشابه فراهم می کند.



شکل ۳: تکنیک‌هایی برای تجزیه و تحلیل احساسات و تشخیص احساسات

۲.۴.۵ تکنیک های تحلیل احساسات

*رویکرد مبتنی بر فرهنگ لغت*این روش یک فرهنگ لغت دارد که در آن به هر کلمه مثبت و منفی یک ارزش احساسی اختصاص داده می شود. سپس از مجموع یا میانگین ارزش‌های احساسی برای محاسبه احساس کل جمله یا سند استفاده می‌شود. با این حال، جورک و همکاران. (۲۰۱۵) رویکرد متفاوتی به نام تابع نرمال سازی را برای محاسبه دقیق ارزش احساسات نسبت به این جمع پایه و تابع میانگین امتحان کرد. رویکرد مبتنی بر فرهنگ لغت و رویکرد مبتنی بر پیکره دو نوع رویکرد مبتنی بر واژگان مبتنی بر واژگان احساسی هستند. به طور کلی، یک فرهنگ لغت کلمات برخی از زبان ها را به صورت سیستمی حفظ می کند، در حالی که یک مجموعه نمونه تصادفی متن در برخی از زبان ها است. معنای دقیق در اینجا در رویکرد مبتنی بر فرهنگ لغت و رویکرد مبتنی بر پیکره اعمال می شود. در رویکرد مبتنی بر فرهنگ لغت،یک فرهنگ لغت از کلمات بذر نگهداری می شود (Schouten و Frasincar۲۰۱۵). برای ایجاد این فرهنگ لغت، اولین مجموعه کوچک از کلمات احساسی، احتمالاً با زمینه های بسیار کوتاه مانند نفی، همراه با برچسب های قطبیت آن جمع آوری شده است (Bern-abé-Moreno et al.۲۰۲۰). سپس فرهنگ لغت با جستجوی مترادف آنها (کلمات با قطبیت یکسان) و متضاد (کلمات با قطبیت مخالف) به روز می شود. دقت تحلیل احساسات از طریق این

رویکرد به الگوریتم بستگی دارد. با این حال، این تکنیک دارای ویژگی دامنه نیست. رویکرد مبتنی بر پیکره محدودیت‌های رویکرد مبتنی بر فرهنگ لغت را با گنجاندن کلمات احساسی خاص دامنه که در آن برچسب قطبیت به کلمه احساس با توجه به زمینه یا دامنه آن اختصاص داده می‌شود، حل می‌کند. این یک رویکرد داده محور است که در آن می توان به کلمات احساسی همراه با زمینه دسترسی داشت. این رویکرد مطمئناً می تواند یک رویکرد مبتنی بر قانون با برخی از تکنیک های تجزیه NLP باشد. بنابراین رویکرد مبتنی بر پیکره تمایل به تعمیم ضعیف دارد، اما می تواند عملکرد عالی را در یک حوزه خاص به دست آورد. از آنجایی که رویکرد مبتنی بر فرهنگ لغت، زمینه پیرامون کلمه احساسی را در نظر نمی گیرد، منجر به کارایی کمتری می شود. بنابراین، چو و همکاران. (۲۰۱۴) به صراحت قطبیت زمینه ای را به کار گرفت تا فرهنگ لغت را در حوزه های متعدد با رویکرد داده محور سازگار کند. آنها گرفتند

استراتژی سه مرحله‌ای: لغت‌نامه‌های مختلف را ادغام کنید، کلماتی را که به طبقه‌بندی کمک نمی‌کنند حذف کنید، و قطبیت را بر اساس یک دامنه خاص تغییر دهید.

SentiWordNet (اصولی و سباستینی۲۰۰۶) و فرهنگ لغت آگاه و منطق کننده احساسات (VADER) (هوتو و گیلبرت)۲۰۱۴) واژگان محبوب در احساسات هستند. جها و همکاران (۲۰۱۸) سعی کرد با ایجاد فرهنگ لغت احساسات به نام فرهنگ لغت آگاه چند دامنه ای هندی (HMDSAD) برای تجزیه و تحلیل احساسات در سطح سند، برنامه واژگان را در حوزه های چندگانه گسترش دهد. از این فرهنگ لغت می توان برای حاشیه نویسی نظرات به مثبت و منفی استفاده کرد. روش پیشنهادی ۲۴ درصد کلمات بیشتری نسبت به واژگان عمومی سنتی هندی (HSWN) Sentiwordnet، یک واژگان خاص دامنه، برچسب‌گذاری کرد. روابط معنایی بین کلمات در واژگان سنتی مورد بررسی قرار نگرفته است، که عملکرد طبقه بندی احساسات را بهبود می بخشد. بر اساس این فرض، ویگاس و همکاران. (۲۰۲۰) واژگان را با گنجاندن اصطلاحات اضافی پس از استفاده از جاسازی کلمات برای کشف مقادیر احساسات برای این کلمات به طور خودکار به روز کرد. این ارزش‌های احساسی از تعبیه‌های واژه‌ای «نزدیک» از واژه‌های موجود در واژگان مشتق شده‌اند.

*رویکرد مبتنی بر یادگیری ماشین*رویکرد دیگری برای تحلیل احساسات وجود دارد که به آن رویکرد یادگیری ماشینی می‌گویند. کل مجموعه داده برای اهداف آموزشی و آزمایشی به دو بخش تقسیم می شود: مجموعه داده آموزشی و مجموعه داده آزمایشی. مجموعه داده آموزشی اطلاعاتی است که برای آموزش مدل با ارائه ویژگی‌های نمونه‌های مختلف یک آیتم استفاده می‌شود. سپس از مجموعه داده آزمایشی استفاده می شود تا ببینیم مدل مجموعه داده آموزشی با چه موفقیتی آموزش داده شده است. به طور کلی، الگوریتم های یادگیری ماشین مورد استفاده برای تجزیه و تحلیل احساسات تحت طبقه بندی نظارت شده قرار می گیرد. انواع مختلفی از الگوریتم‌های مورد نیاز برای طبقه‌بندی احساسات ممکن است عبارتند از Naïve Bayes، ماشین بردار پشتیبان (SVM)، درخت‌های تصمیم‌گیری، و غیره که هر کدام مزایا و معایب خود را دارند. گامون (۲۰۰۴) یک ماشین بردار پشتیبانی را روی ۴۰۸۸۴ بازخورد مشتری جمع آوری شده از نظرسنجی ها اعمال کرد. نویسندگان ترکیب‌های مختلفی از مجموعه ویژگی‌ها را پیاده‌سازی کردند و به دقت ۸۵.۴۷٪ رسیدند. یه و همکاران (۲۰۰۹) با SVM، مدل N-gram و Naïve Bayes روی احساسات و بررسی هفت مقصد محبوب اروپا و ایالات متحده آمریکا کار کرد که از yahoo.com جمع آوری شد. نویسندگان با مدل n-gram به دقت ۸۷.۱۷ درصد دست یافتند. تورفتگی بوچار و همکاران. (۲۰۱۸) واژگانی به نام JOB ۱.۰ ایجاد کرد و مجموعه های خبری به نام SentiNews ۱.۰ را برای تجزیه و تحلیل احساسات در متون اسلوونیایی نامگذاری کرد. JOB ۱.۰ شامل ۲۵۵۲۴ کلمه اصلی است که با مقیاس بندی احساسات از

– ۵ تا ۵ بر اساس مدل AFINN. برای ساخت مجموعه‌ها، داده‌ها از رسانه‌های خبری مختلف وب حذف شدند. سپس، پس از پاکسازی و پیش پردازش داده ها، از حاشیه نویسان خواسته شد تا ۱۰۴۲۷ سند را در مقیاس ۱ تا ۵ حاشیه نویسی کنند که یک به معنای منفی و ۵ به معنای بسیار مثبت است. سپس این اسناد با برچسب‌های مثبت، منفی و خنثی مطابق با میانگین امتیاز مقیاس خاص برچسب‌گذاری شدند. نویسندگان مشاهده کردند که Naïve Bayes در مقایسه با ماشین بردار پشتیبانی (SVM) بهتر عمل کرد. Naive Bayes به امتیاز F۱ بالای ۹۰٪ در طبقه‌بندی باینری و امتیاز F۱ بالای ۶۰٪ برای طبقه‌بندی احساسات سه کلاسه دست یافت. تیواری و همکاران (۲۰۲۰) سه الگوریتم یادگیری ماشین به نام‌های Naive Bayes، SVM و حداکثر آنتروپی را با روش استخراج ویژگی n گرم روی مجموعه داده‌های گوجه‌فرنگی پوسیده پیاده‌سازی کرد. مجموعه داده آموزش و

آزمایش شامل ۱۶۰۰ مرور در هر یک بود. نویسندگان کاهش دقت را با مقادیر بالاتر n در n گرم مشاهده کردند مانند n = چهار، پنج و شش. سومیا و پرامود (۲۰۲۰) ۳۱۸۴ توییت مالایالام را با استفاده از بردارهای ویژگی های مختلف مانند Unigram ،BOW با Sentiwordnet و غیره به نظرات مثبت و منفی طبقه بندی کردند. نویسندگان الگوریتم های یادگیری ماشینی مانند جنگل تصادفی و Naïve Bayes را پیاده سازی کردند و مشاهده کردند که جنگل تصادفی با دقت ۹۵.۶ درصد عملکرد بهتری دارد. Unigram Sentiwordnet با در نظر گرفتن کلمات نفی.

رویکرد *مبتنی بر یادگیری عمیق*در سال‌های اخیر، الگوریتم‌های یادگیری عمیق بر سایر رویکردهای سنتی برای تحلیل احساسات تسلط دارند. این الگوریتم‌ها احساسات یا نظرات متن را بدون انجام مهندسی ویژگی تشخیص می‌دهند. الگوریتم‌های یادگیری عمیق متعددی وجود دارد، یعنی شبکه‌های عصبی مکرر و شبکه‌های عصبی کانولوشنال، که می‌توانند برای تحلیل احساسات اعمال شوند و نتایج دقیق‌تری نسبت به مدل‌های یادگیری ماشین ارائه دهند. این رویکرد باعث می‌شود انسان‌ها از ساختن ویژگی‌های متن به صورت دستی رها شوند، زیرا مدل‌های یادگیری عمیق آن ویژگی‌ها یا الگوها را خودشان استخراج می‌کنند. ژیان و همکاران (۲۰۱۰) از یک مدل مبتنی بر عصبی استفاده کرد
فناوری شبکه برای دسته‌بندی احساسات که شامل ویژگی‌های احساسی، بردارهای وزن ویژگی و پایگاه دانش قبلی است. نویسندگان این مدل را برای بررسی داده های فیلم کرنل به کار بردند. نتایج تجربی این مقاله نشان داد که سطح دقت مدل l در مقایسه با HMM و SVM فوق‌العاده است. پاسوپا و آیوتایا (۲۰۱۹) اعتبارسنجی متقاطع پنج برابری را روی مجموعه داده های داستان کودکان (تایلندی) اجرا کرد و سه مدل یادگیری عمیق به نام های LSTM ،CNN و Bi-LSTM را مقایسه کرد. این مدل‌ها با یا بدون ویژگی‌ها اعمال می‌شوند: برچسب‌گذاری POS (تکنیک پیش پردازش برای شناسایی بخش‌های مختلف گفتار). Thai۲Vec (جاسازی کلمه آموزش دیده از ویکی پدیا تایلندی)؛ sentic (برای درک احساس کلمه). نویسندگان بهترین عملکرد را در مدل CNN با هر سه ویژگی که قبلا اشاره شد مشاهده کردند. همانطور که قبلاً گفته شد، پلتفرم های رسانه های اجتماعی به عنوان منبع مهمی از داده ها در زمینه تحلیل احساسات عمل می کنند. داده‌های جمع‌آوری‌شده از این سایت‌های اجتماعی به دلیل شیوه نوشتن رایگان کاربران، سر و صدای زیادی دارد. بنابراین، آرورا و کانسال (۲۰۱۹) مدلی به نام Conv-char-Emb را پیشنهاد کرد که می تواند مشکل داده های نویزدار را حل کند و از فضای

کوچک حافظه برای جاسازی استفاده کند. برای جاسازی، از شبکه عصبی کانولوشن (CNN) استفاده شده است که از پارامترهای کمتری در نمایش ویژگی استفاده می کند. دشتی پور و همکاران (۲۰۲۰) چارچوب یادگیری عمیق را برای انجام تحلیل احساسات در زبان فارسی پیشنهاد کرد. محققان به این نتیجه رسیدند که شبکه‌های عصبی عمیق مانند LSTM و CNN از الگوریتم‌های یادگیری ماشین موجود در مجموعه داده‌های بررسی هتل و محصول بهتر عمل می‌کنند.

رویکرد *یادگیری انتقالی* و رویکرد *ترکیبی*یادگیری انتقالی نیز بخشی از یادگیری ماشینی است. یک مدل آموزش داده شده بر روی مجموعه داده های بزرگ برای حل یک مشکل می تواند برای سایر مسائل مرتبط اعمال شود. استفاده مجدد از یک مدل از پیش آموزش دیده در حوزه های مرتبط به عنوان نقطه شروع می تواند باعث صرفه جویی در زمان و ایجاد نتایج کارآمدتر شود. ژانگ و همکاران (۲۰۱۲) با مدل سازی مستقیم توزیع بین حوزه های مختلف، یک روش یادگیری نمونه جدید را پیشنهاد کرد. نویسندگان مجموعه داده را طبقه بندی کردند: بررسی محصول آمازون و مجموعه داده های توییتر به احساسات مثبت و منفی. تائو و نیش (۲۰۲۰) گسترش روش های طبقه بندی اخیر در تحلیل احساسات مبتنی بر جنبه را به طبقه بندی چند برچسبی پیشنهاد کرد. نویسندگان همچنین مدل‌های یادگیری انتقالی به نام XLNet و Bert را توسعه دادند و رویکرد پیشنهادی را در مجموعه داده‌های مختلف Yelp ارزیابی کردند. رویکردهای یادگیری عمیق و یادگیری ماشین نتایج خوبی دارند، اما رویکرد ترکیبی می‌تواند نتایج بهتری به همراه داشته باشد زیرا بر محدودیت‌های هر مدل سنتی غلبه می‌کند. ملادنوویچ و همکاران (۲۰۱۶) یک تکنیک کاهش ویژگی، یک چارچوب ترکیبی ساخته شده از واژگان احساسات و wordnet صربی را پیشنهاد کرد. نویسندگان هر دو فرهنگ لغت را با افزودن برخی واژه‌های احساسی صرفی گسترش دادند تا از از دست دادن اطلاعات مهم در حین ریشه‌گیری جلوگیری کنند. آل امرانی و همکاران (۲۰۱۸) مدل هیبریدی ساخته شده خود را مقایسه کردندSVM و مدل جنگل تصادفی، یعنی RFSVM، در بررسی محصول ama-zon. نویسندگان به این نتیجه رسیدند که RFSVM، با سطح دقت ۸۳.۴ درصد، عملکرد بهتری نسبت به SVM با دقت ۸۲.۴ درصد و جنگل تصادفی با دقت ۸۱ درصد به صورت جداگانه در مجموعه داده ۱۰۰۰ بررسی دارد. القریوتی و همکاران (۲۰۲۰) ترکیبی از رویکرد مبتنی بر قانون و واژگان دامنه را برای تشخیص احساسات در سطح جنبه برای درک نظرات مردم در مورد برنامه‌های هوشمند دولتی پیشنهاد کرد. نویسندگان به این نتیجه رسیدند

که تکنیک پیشنهادی ۵ درصد از سایر مدل‌های پایه مبتنی بـر واژگـان برتـری دارد. ری و چاکرابـارتی (۲۰۲۰) رویکرد مبتنی بر قانون را برای استخراج جنبه ها با یک مدل یادگیری عمیق ۷ لایه CNN بـرای برچسب گذاری هر جنبه ترکیب کرد. مدل ترکیبی دقت ۸۷ درصـدی را بـه دسـت آورد، در حـالی کـه مدل‌های منفرد ۷۵ درصد دقت مبتنی بر قانون و ۸۰ درصد دقت با مدل CNN داشتند.

۲.۴.۶ تکنیک های تشخیص احساسات

*رویکرد مبتنی بر واژگان*رویکرد مبتنی بر واژگان یک رویکرد جستجوی مبتنی بر کلمه کلیدی است که به جستجوی کلمات کلیدی احساسات اختصاص داده شده به بـرخی از حـالات روانی می‌پـردازد (رابیـا و همکـاران۲۰۱۷). واژگـان محبـوب بـرای تشـخیص احساسـات عبارتنـد از Word-Net-Affect Strapparava et al.۲۰۰۴)و NRC واژگـان کلمـه-احسـاس (محمـد و تـورنی۲۰۱۳). WordNet -Affect یک فرم توسعه یافته از WordNet است که از کلمـات عـاطفی تشـکیل شده است که با برچسب های احساسات حاشیه نویسی شده انـد. واژگـان NRC شـامل ۱۴۱۸۲ کلمـه است که هر کدام به یک احساس خاص و دو احساس اختصاص دارد. این واژگان، واژگان دسـته بنـدی هستند که هر کلمه را با یک حالت احساسی برای طبقه بندی احساسات برچسب گذاری می کنند. با این حال، با نادیده گرفتن شدت احساسات، این واژگـان سـنتی کمـتر آموزنـده و کمـتر سـازگار می شـوند. بنابراین، لی و همکاران. (۲۰۲۱) یک استراتژی موثر برای به دست آوردن توزیع هیجان در سـطح کلمـه پیشـنهاد کـرد تـا بـا ادغـام یـک فرهنـگ لغت بعـدی بـه نـام NRC-Valence arosal dominance، احساسات با شدت بـه کلمـات احساسـی اختصـاص یابـد. EmoSenticNet (پوریا و همکاران۲۰۱۴) همچنین شامل تعداد زیادی است که به برچسب هـای کیفی و کمی اختصـاص داده شده است. به طور کلی، محققان واژگان خود را تولید می کننـد و مسـتقیماً آنهـا را بـرای تجزیـه و تحلیل احساسات به کار می برند، اما واژگان می توانند برای اهداف استخراج ویژگی نیز اسـتفاده شـوند. عبداوی و همکاران (۲۰۱۷) از استفاده از ابزارهای ترجمه آنلاین برای ایجـاد واژگـان فرانسـوی بـه نـام FEEL (فرانسوی واژگان عاطفی گسترش یافته) اسـتفاده کـرد کـه شـامل بیش از ۱۴۰۰۰ کلمـه بـا دو قطبیت و برچسب احساسات است. این واژگان توسـط افـزایش تعـداد کلمـات در واژگـان احساسـات NRC و ترجمه نیمه خودکار با استفاده از شش مترجم آنلاین. آن مدخل‌هایی که از حداقل سه مـترجم

به‌دست می‌آمدند، از پیش تأیید شده در نظر گرفته شدند و سپس توسط مترجم دستی تأیید شدند. بندکاوی و همکاران (۲۰۱۷) یک واژگان خاص دامنه را برای فرآیند استخراج ویژگی در تحلیل احساسات به کار برد. نویسندگان به این نتیجه رسیدند که ویژگی های مشتق شده از واژگان پیشنهادی آنها از سایر ویژگی های پایه بهتر عمل می کند. براون و همکاران (۲۰۲۱) یک مجموعه چند زبانه به نام MEmoFC که مخفف عبارت Multilingual Emotional Football Corpus است، ساخته است که شامل گزارش‌های فوتبال از وب‌سایت‌های انگلیسی، هلندی و آلمانی و آمار مسابقات خزیده‌شده از Goal.com است. مجموعه با ایجاد دو جدول فراداده ایجاد شد: یکی توضیح جزئیات یک مسابقه مانند تاریخ، مکان، تیم های شرکت و غیره، و جدول دوم شامل اختصارات باشگاه های فوتبال است. نویسندگان مجموعه را با رویکردهای مختلف برای دانستن تأثیر گزارش ها بر نتایج بازی نشان دادند.

*تکنیک های مبتنی بر یادگیری ماشین*تشخیص یا طبقه‌بندی احساسات ممکن است به انواع مختلفی از مدل‌های یادگیری ماشین نیاز داشته باشد، مانند Naïve Bayes، ماشین بردار پشتیبان، درخت‌های تصمیم، و غیره. (۲۰۱۷) احساسات را از متون چند زبانه جمع آوری شده از سه حوزه مختلف استخراج کرد. نویسندگان از یک رویکرد جدید به نام خلاصه سایت غنی برای جمع‌آوری داده‌ها استفاده کردند و از الگوریتم‌های یادگیری ماشینی SVM و Naïve Bayes برای طبقه‌بندی احساسات متن توییتر استفاده کردند. نتایج نشان داد که سطح دقت ۷۱.۴٪ با الگوریتم Naïve Bayes به دست آمد. حسن و همکاران (۲۰۱۹) الگوریتم های یادگیری ماشینی مانند SVM، Naïve Bayes و درخت های تصمیم را برای شناسایی احساسات در پیام های متنی ارزیابی کرد. این کار به دو زیرکار تقسیم می شود: وظیفه ۱ شامل مجموعه ای از مجموعه داده از توییتر و برچسب زدن خودکار مجموعه داده با استفاده از هشتگ ها و آموزش مدل است. ۲ Task در حال توسعه EmotexStream دو مرحله‌ای است که توییت‌های بدون احساس را در مرحله اول جدا می‌کند و با استفاده از مدل‌های آموزش‌دیده در task۱، احساسات را در متن شناسایی می‌کند. نویسندگان دقت ۹۰ درصدی را در طبقه بندی احساسات مشاهده کردند. اصغر و همکاران (۲۰۱۹) با هدف اعمال چندین مدل یادگیری ماشین در مجموعه داده ISEAR برای یافتن بهترین طبقه‌بندی‌کننده. آنها دریافتند که مدل رگرسیون لجستیک بهتر از سایر طبقه بندی کننده ها با ارزش یادآوری ۸۳ درصد عمل می کند.

یادگیری عمیق و تکنیک ترکیبی حوزه یادگیری عمیق بخشی از یادگیری ماشینی است که اطلاعات یا سیگنال ها را به همان روشی که مغز انسان انجام می دهد پردازش می کند. مدل های یادگیری عمیق حاوی چندین لایه نورون هستند. هزاران نورون به یکدیگر متصل هستند، که سرعت پردازش را به صورت موازی افزایش می دهد. چت ترجی و همکاران (۲۰۱۹) مدلی به نام تشخیص احساسات و احساسات معنایی (SSBED) با تغذیه احساسات و بازنمایی های معنایی به ترتیب به دو لایه LSTM ایجاد کرد. این نمایش‌ها سپس به هم متصل می‌شوند و سپس برای طبقه‌بندی به شبکه مش منتقل می‌شوند. رمان رویکرد مبتنی بر احتمال وجود احساسات متعدد در جمله است و از بازنمایی معنایی و احساسی برای طبقه‌بندی بهتر احساسات استفاده می‌کند. نتایج بر روی مجموعه داده‌های ساخته‌شده خودشان با جفت‌های مکالمه توییت ارزیابی می‌شوند و مدل آنها با سایر مدل‌های پایه مقایسه می‌شود. خو و همکاران (۲۰۲۰) با استفاده از مدل‌های دو ترکیبی به‌ترتیب با نام‌های حافظه کوتاه‌مدت متحرک-طولانی سه بعدی (۳DCLS) و CNN-RNN احساسات را از ویدیو و متن استخراج کرد. در همان زمان، نویسندگان SVM را برای طبقه‌بندی احساسات مبتنی بر صوتی پیاده‌سازی کردند. نویسندگان نتایج را با ادغام ویژگی‌های صوتی و تصویری در سطح ویژگی با تکنیک MKL fusion و ترکیب بیشتر نتایج آن با نتایج طبقه‌بندی احساسات مبتنی بر متن به نتیجه رسیدند. این روش دقت بهتری نسبت به سایر تکنیک‌های همجوشی چندوجهی ارائه می‌کند و قصد دارد احساسات بررسی‌های دارویی نوشته شده توسط بیماران در پلتفرم‌های رسانه‌های اجتماعی را تحلیل کند. بصیری و همکاران (۲۰۲۰) دو مدل را با استفاده از تئوری تصمیم گیری سه طرفه پیشنهاد کرد. مدل اول ادغام سه‌طرفه یک مدل یادگیری عمیق با روش یادگیری سنتی (۳W۱DT) است، در حالی که مدل دیگر تلفیقی سه‌طرفه از سه مدل یادگیری عمیق با روش یادگیری معمولی (۳W۳DT) است. نتایج به دست آمده با استفاده از مجموعه داده Drugs.com نشان داد که هر دو چارچوب بهتر از تکنیک‌های یادگیری عمیق سنتی عمل می‌کنند. علاوه بر این، عملکرد مدل فیوژن اول در مقایسه با مدل دوم از نظر دقت و متریک F۱ بسیار بهتر بود. در روزهای اخیر، پلتفرم‌های رسانه‌های اجتماعی مملو از پست‌های مرتبط با کووید-۱۹ هستند. سینگ و همکاران (۲۰۲۱) تجزیه و تحلیل تشخیص احساسات را روی توییت‌های کووید-۱۹ جمع‌آوری‌شده از کل جهان و هند تنها با مدل‌های رمزگذار دوطرفه از ترانسفورماتورها (BERT) روی مجموعه داده‌های توییتر اعمال کرد و تقریباً ۹۴ درصد دقت را به دست آورد.

رویکرد *یادگیری انتقالی*در رویکردهای سنتی، فرض رایج این است که مجموعه داده از یک حوزه است. با این حال، زمانی که دامنه تغییر می کند، نیاز به یک مدل جدید وجود دارد. رویکرد یادگیری انتقال به شما امکان می دهد از مدل های از پیش آموزش دیده موجود در حوزه هدف مجددا استفاده کنید. برای مثال احمد و همکاران. (۲۰۲۰) از تکنیک یادگیری انتقالی به دلیل کمبود منابع برای تشخیص احساسات در زبان هندی استفاده کرد. محققان از قبل آموزش دیدند

مدل بر روی دو مجموعه داده انگلیسی مختلف: SemEval-۲۰۱۸، تجزیه و تحلیل احساسات، و یک مجموعه داده هندی با برچسب‌های مثبت، خنثی، درگیری و منفی. آنها امتیاز ۰.۵۳ f۱ را با استفاده از یادگیری انتقال و ۰.۴۷ با استفاده از مدل های پایه CNN و Bi-LSTM با جاسازی کلمه متقابل زبانی به دست آوردند. هزاریکا و همکاران (۲۰۲۰) یک مدل TL-ERC ایجاد کرد که در آن مدل از قبل بر روی مکالمات چند نوبتی منبع آموزش داده شد و سپس بر روی وظیفه طبقه بندی احساسات در پیام های رد و بدل شده منتقل شد. نویسندگان بر موضوعاتی مانند کمبود داده های برچسب گذاری شده در مکالمات چندگانه با چارچوب مبتنی بر یادگیری انتقال استقرایی تأکید کردند.

اکثر محققان مدل‌ها را با ترکیب تکنیک‌های یادگیری ماشین و یادگیری عمیق با تکنیک‌های مختلف استخراج ویژگی پیاده‌سازی کردند. بیشتر مجموعه داده ها به زبان انگلیسی در دسترس هستند. با این حال، برخی از محققان مجموعه داده های زبان منطقه ای خود را ساختند. به عنوان مثال، ساسیذر و همکاران. (۲۰۲۰) مجموعه داده کد هندی-انگلیسی را با سه احساس اصلی ایجاد کرد: خوشحال، غمگین و عصبانی، و مشاهده شد که CNN-BILSTM عملکرد بهتری نسبت به دیگران دارد.

۳ روش پیشنهادی

در این روش ما تمام مراحلی را که برای پردازش داده‌های دیتاست استفاده شده است پیاده‌سازی میکنیم. مراحلی که ما استفاده کردیم عبارتند از:

برای پیاده سازی یک مدل روی یک مجموعه داده ما قدم های زیر را طی خواهیم کرد:

۱. شناختن داده ها : در این بخش ما تعدادی بررسی آماری روی مجموعه داده انجام خواهیم داد که داده هایی که قرار است با آنها کار کنیم را بهتر شناخته و به یک دید کلی نسبت به داده ها برسیم.

۲. پیش پردازش داده ها : در بخش پیش پردازش ما از تکنیک هایی استفاده میکنیم که بتوانیم به ویژگی های مناسبی که مد نظر ما هستند و تاثیر گذاری مستقیم در خروجی دارند برسیم و داده های اضافی که تاثیری در خروجی ندارند و یا ما نمیخواهیم تاثیر آنها را مورد بحث و بررسی قرار دهیم حذف کنیم. همچنین داده ها را به فرم مناسبی در آورده که برای مدل هایی که میخواهیم پیاده سازی کنیم قابل پردازش باشند.

۳. تقسیم کردن داده ها : از آنجایی که بعد از ساختن مدل ما نیاز به ارزیابی مدل و بحث و بررسی در مورد دقت مدل خواهیم داشت میبایست داده ها را به دو بخش داده یادگیری و تست تقسیم کنیم که

بتوان بعد از خوراندن داده های یادگیری به مدل و ساخته شدن مدل با استفاده از داده های تست دقت مدل را مورد ارزیابی قرار داد

۴. ساختن مدل : در نهایت میتوان مدل خود را با الگوریتم های متفاوت ساخت و مدل های ساخته شده را مورد بحث و بررسی قرار داد و دقت هرکدام را سنجید. انتخاب مدل به عوامل متعددی وابسته است که مهم ترین آن ها پراکندگی آماری داده ها و تعداد کلاس های موجود هستند.

۳.۱ شناختن داده‌ها

شناختن داده‌ها برای کار کردن با آن‌ها از اهمیت بالایی است در ادامه به چندی از آن‌ها اشاره میکنیم:

۱. **تضمین کیفیت داده**: هنگام کار با وظایف NLP، باید از کیفیت داده‌هایی که استفاده می‌کنید مطمئن باشید. این شامل درک منبع داده ها، فرآیند جمع آوری و سوگیری های بالقوه است. به عنوان مثال، اگر در حال ساخت یک مدل تجزیه و تحلیل احساسات هستید، دانستن اینکه آیا منبع داده توییتر است یا مقاله های خبری رسمی می تواند تا حد زیادی بر زبان و عبارات احساسات در مجموعه داده تأثیر بگذارد.

۲. **تشخیص سوگیری و کاهش**: درک مجموعه داده ها به تشخیص سوگیری ها کمک می کند. مدل‌های NLP می‌توانند ناخواسته تعصب‌های موجود در داده‌های آموزشی را یاد بگیرند و منتشر کنند. دانستن منبع و زمینه مجموعه داده شما را قادر می سازد تا چنین سوگیری ها را شناسایی و کاهش دهید و از عدالت و برابری در برنامه های NLP خود اطمینان حاصل کنید.

۳. **حجم و تنوع داده ها**: وظایف مختلف NLP به مقادیر متفاوتی از داده نیاز دارند. دانستن اندازه و تنوع مجموعه داده ها برای انتخاب الگوریتم ها و مدل های مناسب بسیار مهم است. برای مثال، مدل‌های یادگیری عمیق اغلب به مجموعه داده‌های بزرگ برای آموزش مؤثر نیاز دارند، در حالی که مدل‌های ساده‌تر ممکن است برای مجموعه داده‌های کوچک‌تر کافی باشند.

۴. **برچسب‌گذاری و حاشیه‌نویسی داده‌ها**: در بسیاری از وظایف NLP، مانند طبقه‌بندی متن یا شناسایی موجودیت نام‌گذاری شده، برچسب‌گذاری دستی یا حاشیه‌نویسی داده‌ها ضروری

است. درک مجموعه داده‌ها می‌تواند به تعریف دستورالعمل‌های برچسب‌گذاری واضح و اطمینان از ثبات در حاشیه‌نویسی کمک کند، که برای آموزش مدل ضروری است.

۵. **استراتژی های پیش پردازش داده ها**: بسته به مجموعه داده، ممکن است به استراتژی های پیش پردازش متفاوتی نیاز داشته باشید. به عنوان مثال، متن از رسانه های اجتماعی ممکن است در مقایسه با مقالات دانشگاهی نیاز به تمیز کردن و کاهش نویز گسترده تری داشته باشد. دانستن ماهیت داده‌ها، تصمیم‌گیری در مورد تکنیک‌هایی مانند ریشه‌یابی، واژه‌سازی یا مدیریت کاراکترهای خاص را تعیین می‌کند.

۶. **دانش اختصاصی دامنه**: وظایف NLP اغلب شامل اصطلاحات و زمینه خاص دامنه است. دانستن دامنه مجموعه داده (به عنوان مثال، پزشکی، حقوقی، مالی) برای استخراج موثر ویژگی، سفارشی سازی مدل، و درک ظرافت های متن ضروری است.

۷. **تنوع زبان**: مناطق و جوامع مختلف ممکن است زبان را به طور متفاوتی استفاده کنند. تشخیص منشأ مجموعه داده و زبان های استفاده شده در آن، به ویژه برای کارهایی مانند شناسایی زبان یا تشخیص گویش، مهم است.

۸. **تشخیص ناهنجاری**: در برخی از برنامه های NLP، مانند تشخیص تقلب یا فیلتر کردن هرزنامه، دانستن ویژگی های داده های عادی و غیرعادی بسیار مهم است. درک مجموعه داده به شناسایی آنچه رفتار یا زبان عادی را تشکیل می دهد و در نتیجه، تشخیص ناهنجاری ها کمک می کند.

۹. **افزایش داده ها**: تکنیک های افزایش داده ها، مانند جایگزینی مترادف یا ترجمه متن، زمانی که درک خوبی از محتوا و زمینه مجموعه داده داشته باشید، می توانند به طور موثرتری اعمال شوند.

در اصل، دانستن مجموعه داده مانند درک مواد خامی است که در NLP با آن کار می کنید. این شما را در تصمیم گیری آگاهانه در هر مرحله از خط لوله NLP، از جمع آوری داده ها و پیش پردازش گرفته تا آموزش و ارزیابی مدل، راهنمایی می کند. این درک به شما اجازه می دهد تا رویکرد خود را با ویژگی ها و چالش های خاص ناشی از داده ها تطبیق دهید و در نهایت منجر به نتایج و برنامه های کاربردی بهتر NLP می شود.

مجموعه داده‌ای که ما با آن کار خواهیم کرد مشخصات زیر را دارد:

- ستون اول user ID کاربر
- ستون دوم متن خام توییت (۴۰.۰۰۰)
- ستون سوم احساس(۱۳ مدل)

این مجموعه داده دامنه عمومی از پلتفرم data.world جمع آوری شده است. با تشکر، data.world برای انتشار آن.

۳.۲ پیش پردازش داده‌ها

برای آماده سازی داده‌ها ما کار های زیر را انجام دادیم:

۱. cleaning
۲. tokenization
۳. removing stopwords
۴. lemmatization
۵. stemming

کدی که برای این مرحله زدیم به شرح زیر است:

۳.۲.۱ حذف کردن url

حذف URL ها از داده های متنی، به ویژه در زمینه NLP برای توییت ها یا رسانه های اجتماعی، به چند دلیل یک مرحله پیش پردازش رایج است:

- **کاهش نویز**: URL های موجود در توییت ها اغلب حاوی نویز زیادی از جمله کاراکترها، نمادها و اعداد تصادفی هستند. حذف آنها به کاهش نویز در داده های متن کمک می کند و آن را تمیزتر و مناسب تر برای تجزیه و تحلیل می کند.

- **یکنواختی**: حذف URL ها تضمین می کند که داده های متنی سازگارتر هستند. از آنجایی که URL ها می توانند از نظر طول و محتوا به طور قابل توجهی متفاوت باشند، حضور آنها می تواند تنوعی را در مجموعه داده ایجاد کند که به وظیفه NLP مرتبط نیست.
- **تمرکز بر محتوای متنی**: وظایف NLP معمولاً بر تجزیه و تحلیل و درک محتوای متنی داده ها متمرکز است. URL ها که آدرس های وب هستند، اغلب حاوی اطلاعات معناداری برای تجزیه و تحلیل خود متن نیستند. حذف URL ها به تمرکز روی کلمات و عباراتی که پیام یا احساس موجود در توییت را منتقل می کنند کمک می کند.
- **کاهش ابعاد**: URL ها می توانند بسیار طولانی باشند، و ممکن است اغلب در توییت ها، به خصوص در ریتوییت ها یا هنگام اشتراک گذاری لینک ها، ظاهر شوند. گنجاندن URL ها در داده های متنی می تواند ابعاد مجموعه داده را به میزان قابل توجهی افزایش دهد. حذف آنها ابعاد را کاهش می دهد و آن را برای کارهای NLP قابل مدیریت و کارآمدتر می کند.
- **ثبات در توکن سازی**: زمانی که متن را نشانه گذاری می کنید (آن را به کلمات یا نشانه ها تقسیم می کنید)، URL ها اغلب به چندین توکن تقسیم می شوند که هر کدام شامل بخش هایی از URL است. حذف نشانی‌های اینترنتی تضمین می‌کند که توکن‌سازی به جای تکه‌هایی از URLها، به کلمات معنادار منجر می‌شود.
- **حریم خصوصی و امنیت**: URL ها گاهی اوقات می توانند به وب سایت ها یا منابع خارجی منجر شوند که ممکن است پیامدهای حفظ حریم خصوصی داشته باشند. با حذف URL ها، می توانید از حریم خصوصی کاربران محافظت کنید و اطمینان حاصل کنید که اطلاعات حساس به طور ناخواسته در معرض دید قرار نمی گیرند یا به آنها دسترسی پیدا نمی شود.
- **عملکرد مدل بهبود یافته**: مدل‌های NLP اغلب زمانی که بر روی داده‌های متنی تمیز و از پیش پردازش شده آموزش می‌بینند، عملکرد بهتری دارند. حذف URL ها یک مرحله در خط لوله پیش پردازش است که می تواند با کاهش نویز و اطلاعات نامربوط به بهبود عملکرد مدل کمک کند.
- **افزایش خوانایی**: برای خوانندگان انسانی یا تحلیلگرانی که متن را بررسی می کنند، URL ها چندان آموزنده نیستند و می توانند جریان فهم را مختل کنند. حذف آنها خوانایی متن را افزایش می دهد.

به طور خلاصه، حذف URL ها از داده های متنی در NLP برای توییت ها یک روش معمول برای تمیز کردن و آماده سازی داده ها برای تجزیه و تحلیل، بهبود عملکرد مدل و تمرکز بر محتوای متنی مربوطه توییت ها است. این بخشی از مجموعه گسترده تری از مراحل پیش پردازش متن است که با هدف استخراج اطلاعات معنی دار از متن و در عین حال کاهش نویز و اطلاعات نامربوط انجام می شود.

ما با استفاده از این قطعه کد لینک ها را از مجموعه داده خود حذف کردیم:

```
# Function to remove URLs from text
def remove_urls(text):
    url_pattern = re.compile(r'https?://\S+|www\.\S+')
    return url_pattern.sub('', text)
```

۳.۲.۲ حذف mentions

اشاره در توییت‌ها که به نام‌های کاربری یا دسته‌های توییتر نیز شناخته می‌شوند، اشاره‌ای به سایر کاربران توییتر هستند. قبل از آنها علامت "@" و پس از آن نام کاربری وجود دارد. در اینجا توضیح مختصری در مورد اینکه چرا اشاره ها معمولاً در طول پیش پردازش متن برای وظایف NLP شامل توییت ها به روشی خاص برخورد می شوند آورده شده است:

۱. **شناسایی تعاملات کاربر**: اشاره ها برای درک نحوه تعامل کاربران با یکدیگر در توییتر بسیار مهم هستند. آنها نشان می دهند که یک کاربر مستقیماً به کاربر دیگری آدرس می دهد یا به او ارجاع می دهد. تجزیه و تحلیل اشاره ها می تواند به مطالعه تعاملات اجتماعی، احساسات و نفوذ در جامعه توییتر کمک کند.

۲. **حذف اطلاعات خاص کاربر**: در برخی از کارهای NLP، ممکن است بخواهید برای محافظت از حریم خصوصی کاربر یا تمرکز صرفاً بر محتوای متنی توییت ها، اشاره ها را حذف کنید. از آنجایی که اشاره ها شامل نام‌های کاربری خاص هستند، حذف آنها می‌تواند به حفظ ناشناس بودن و تعمیم تحلیل کمک کند.

۳. **کاهش ابعاد**: مشابه URL ها، اشاره می تواند اغلب در توییت ها، به خصوص در مکالمات، پاسخ ها، یا ریتوییت ها اتفاق بیفتد. گنجاندن نام‌های موجود در داده‌های متنی می‌تواند ابعاد مجموعه داده را به میزان قابل توجهی افزایش دهد. حذف نام‌ها ابعاد را کاهش می‌دهد و می‌تواند منجر به مدل‌سازی کارآمدتر شود.

۴. **ثبات توکن سازی**: هنگام توکن کردن متن، اشاره ها اغلب به نشانه های جداگانه تقسیم می شوند، یکی برای نماد "@" و دیگری برای نام کاربری. حذف نام‌ها تضمین می‌کند که توکن‌سازی به‌جای تکه‌هایی از نام‌های کاربری منجر به کلمات معنادار می‌شود.

۵. **ساده کردن تجزیه و تحلیل متن**: در برخی از وظایف NLP، ممکن است علاقه ای به تجزیه و تحلیل تعاملات بین کاربران خاص یا مطالعه اشاره ها به عنوان تمرکز اصلی نداشته باشید. حذف اشاره ها با تمرکز بر محتوا و احساسات توییت بدون در نظر گرفتن تعاملات کاربر، تجزیه و تحلیل متن را ساده می کند.

۶. **استخراج محتوا**: برای تجزیه و تحلیل احساسات یا مدل‌سازی موضوع، اشاره ها ممکن است نویز در نظر گرفته شوند زیرا معمولاً حاوی احساسات یا اطلاعات موضوعی نیستند. حذف نام‌ها به شما امکان می‌دهد تا محتوای اصلی توییت را برای تجزیه و تحلیل استخراج کنید.

۷. **احترام به داده های کاربر**: حذف نام ها می تواند راهی برای احترام به حریم خصوصی و مالکیت داده های کاربران توییتر باشد. در تحقیقات یا برنامه های کاربردی، مهم است که داده های کاربر را مسئولانه و مطابق با دستورالعمل های حفظ حریم خصوصی مدیریت کنید.

قطعه کدی که برای حذف mentions بود به شرح زیر است:

```
# Function to remove mentions from text
def remove_mentions(text):
    mention_pattern = re.compile(r'@\w+')
    return mention_pattern.sub('', text)
```

۳.۲.۳ تمیز کردن متن

پاکسازی متن یک مرحله پیش پردازش ضروری در پردازش زبان طبیعی (NLP) است که شامل فرآیند تهیه داده های متن خام برای تجزیه و تحلیل است. در اینجا توضیحی در مورد اینکه چرا پاکسازی متن بسیار مهم است و مراحل مربوطه آورده شده است:

اهمیت پاکسازی متن در NLP:

۱. **کاهش نویز**: داده های متنی اغلب حاوی نویز به شکل کاراکترهای خاص، علائم نگارشی، تگ های HTML و نمادهای نامربوط هستند. پاک کردن متن این عناصر را حذف می کند و داده ها را برای تجزیه و تحلیل مناسب تر می کند.

۲. **سازگاری**: تمیز کردن تضمین می کند که داده های متنی سازگار و استاندارد هستند. تغییرات در حروف بزرگ، املا و قالب بندی را می توان کاهش داد، که منجر به یک مجموعه داده یکنواخت تر می شود.

۳. **توکن‌سازی بهبودیافته**: توکن‌سازی، فرآیند تقسیم متن به کلمات یا نشانه‌ها، بر روی متن تمیز مؤثرتر است. متن پاک شده منجر به توکن سازی دقیق تری می شود که برای کارهای پایین دستی NLP حیاتی است.

۴. **حذف کلمات توقف**: تمیز کردن اغلب شامل حذف کلمات توقف (کلمات رایج مانند "و"، "the"، "است") است که معنای زیادی ندارند. این مرحله ابعاد داده ها را کاهش می دهد و روی کلمات معنی دارتر تمرکز می کند.

۵. **خوانایی پیشرفته**: خواندن و درک متن پاک شده آسان تر است، هم برای انسان و هم برای مدل های NLP. درک بهتر محتوای متن را تسهیل می کند.

۶. **کاهش ابعاد**: با حذف کاراکترها و نمادهای غیر ضروری، تمیز کردن می تواند ابعاد داده ها را به میزان قابل توجهی کاهش دهد. این امر به ویژه هنگام کار با مجموعه داده های بزرگ مهم است.

۷. **عملکرد مدل بهبود یافته**: مدل های NLP اغلب زمانی که بر روی متن تمیز و از قبل پردازش شده آموزش می بینند بهتر عمل می کنند. داده‌های پاک ابهام و نویز را کاهش می‌دهد که می‌تواند منجر به پیش‌بینی دقیق‌تر مدل شود.

مراحل پاکسازی متن:

تمیز کردن متن معمولاً شامل یک سری مراحل است که ممکن است شامل موارد زیر باشد:

۱. **حروف کوچک**: تبدیل تمام متن به حروف کوچک برای اطمینان از سازگاری و کاهش تأثیر حروف بزرگ.
۲. **حذف کاراکترهای خاص**: حذف کاراکترهای خاص، علائم نقطه گذاری و نمادهایی که حاوی اطلاعات معنی دار نیستند.
۳. **مدیریت اعداد**: بسته به کار، می توانید اعداد را در متن نگه دارید، حذف کنید یا جایگزین کنید.
۴. **حذف برچسب های HTML**: هنگام کار با داده های وب، حذف برچسب های HTML برای استخراج فقط محتوای متنی بسیار مهم است.
۵. **ریشه یا Lemmatization**: کاهش کلمات به شکل پایه یا ریشه آنها. ساقه زنی تهاجمی تر است، در حالی که lemmatization دقیق تر است.
۶. **Stopword Removal**: حذف کلمات توقف متداول که کمک زیادی به معنای متن ندارند.
۷. **بررسی املا**: تصحیح اشتباهات املایی، در صورت وجود.
۸. **Tokenization**: تقسیم متن به کلمات یا نشانه های جداگانه برای تجزیه و تحلیل بیشتر.
۹. **تمیز کردن سفارشی**: بسته به مجموعه داده و کار خاص، مراحل تمیز کردن اضافی ممکن است لازم باشد، مانند مدیریت ایموجی ها، URL ها، یا ذکرها.

ما در این پروژه از قطعه کد زیر برای تمیز کردن متن استفاده کردیم:

```
# Function to cleaning the text
def clean_text(text):
    # Remove special characters, hashtags, and mentions
    cleaned_text = re.sub(r"[^a-zA-Z0-9\s]", "", text)
    #cleaned_text = re.sub(r"#\w+", "", cleaned_text)
    cleaned_text = re.sub(r"@\w+", "", cleaned_text)
```

```
    # Convert text to lowercase
    cleaned_text = cleaned_text.lower()

    # Remove extra whitespace
    cleaned_text = re.sub(r"\s+", " ", cleaned_text).strip()

    return cleaned_text

# Function to tokenize
def tokenize_text(text):
    # Tokenize the text into individual words/tokens
    tokens = word_tokenize(text)
    return tokens

# Function to remove stop words from text
def remove_stopwords(tokens):
    stop_words = set(stopwords.words('english'))
    filtered_tokens = [token for token in tokens if token.lower() not in
stop_words]
    return filtered_tokens

# Function to remove punctuations and numbers
def remove_punctuations_and_numbers(text):
    # Remove punctuation marks
    text = text.translate(str.maketrans('', '', string.punctuation))

    # Remove numbers
    text = re.sub(r'\d+', '', text)

    return text

# Lemmatizer Function
```

```
def perform_lemmatization(tokens):
    lemmatizer = WordNetLemmatizer()
    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in tokens]
    return lemmatized_tokens

# Stemmer Function
def perform_stemming(tokens):
    stemmer = PorterStemmer()
    stemmed_tokens = [stemmer.stem(token) for token in tokens]
    return stemmed_tokens
```

در مرحله بعدی ما داده‌ها را در فایل جداگانه ذخیره میکنیم که در مرحله بعدی از فایل جدید استفاده کنیم.

```
# Save to pre_proccessed.csv
df.to_csv('pre_proccessed.csv', index=False)
```

۳.۳ تقسیم داده‌ها

تقسیم یک مجموعه داده به زیرمجموعه های آموزشی و آزمایشی، به چند دلیل مهم، یک عمل اساسی در یادگیری ماشین، از جمله پردازش زبان طبیعی (NLP) است:

۱. **ارزیابی مدل**: هدف اصلی از تقسیم مجموعه داده، ارزیابی عملکرد یک مدل یادگیری ماشین است. با رزرو بخشی از داده ها برای آزمایش (مجموعه تست)، می توانید ارزیابی کنید که مدل شما چقدر به نمونه های جدید و دیده نشده تعمیم می یابد. بدون یک مجموعه آزمایشی جداگانه، راه قابل اعتمادی برای اندازه گیری عملکرد مدل خود در داده های دنیای واقعی نخواهید داشت.

۲. **تشخیص بیش‌برازش**: تقسیم داده‌ها به تشخیص اضافه برازش کمک می‌کند، که زمانی اتفاق می‌افتد که یک مدل داده‌های آموزشی را خیلی خوب یاد می‌گیرد اما نمی‌تواند به داده‌های جدید تعمیم یابد. اگر یک مدل در داده‌های آموزشی عملکرد فوق‌العاده‌ای داشته باشد اما در داده‌های آزمایشی ضعیف عمل کند، این نشانه بیش از حد برازش است. این به شما اطلاع می دهد که

مدل ممکن است خیلی پیچیده باشد یا داده های آموزشی را به جای یادگیری از آن به خاطر بسپارد.

۳. **تنظیم فراپارامتر**: در طول توسعه مدل‌های یادگیری ماشین، اغلب نیاز به تنظیم دقیق فراپارامترها (به عنوان مثال، نرخ یادگیری، نقاط قوت منظم‌سازی) دارید تا به بهترین عملکرد برسید. شما از مجموعه تست برای ارزیابی تاثیر این تغییرات فراپارامتر بر عملکرد مدل استفاده می‌کنید، بدون اینکه بایاس از داده‌های آموزشی وارد کنید.

۴. **انتخاب مدل**: در برخی موارد، ممکن است بخواهید چندین مدل را با هم مقایسه کنید تا مشخص کنید کدام یک در داده های دیده نشده بهترین عملکرد را دارد. مجموعه تست به شما امکان می دهد این مقایسه را به صورت عینی انجام دهید و مناسب ترین مدل را برای کار NLP خود انتخاب کنید.

۵. **جلوگیری از نشت داده ها**: نگه داشتن یک مجموعه آزمایشی مجزا به شما کمک می کند تا اطمینان حاصل کنید که در طول آموزش مدل، سهواً از هیچ اطلاعاتی از داده های آزمایش استفاده نکنید. نشت داده ها می تواند منجر به تخمین های عملکرد بیش از حد خوش بینانه شود، و باعث می شود مدل شما بهتر از آنچه هست ظاهر شود.

۶. **شبیه سازی دنیای واقعی**: مجموعه تست نحوه عملکرد مدل شما را در یک سناریوی واقعی که در آن با داده های جدید و دیده نشده مواجه می شود، شبیه سازی می کند. این برای ارزیابی اینکه آیا مدل NLP شما برای استقرار آماده است یا خیر حیاتی است.

۷. **اعتبارسنجی متقاطع**: علاوه بر یک تقسیم آزمایش ساده قطار، می‌توانید تکنیک‌های پیشرفته‌تری مانند اعتبارسنجی متقاطع k-fold را انجام دهید. اعتبار سنجی متقابل داده ها را به زیر مجموعه های متعدد تقسیم می کند و به شما این امکان را می دهد که مدل خود را چندین بار روی ترکیب های مختلف مجموعه های آموزشی و اعتبارسنجی آموزش و آزمایش کنید. این یک تخمین قوی تر از عملکرد مدل ارائه می دهد.

۸. **ارزیابی تعمیم**: تقسیم داده ها به شما امکان می دهد توانایی مدل خود را برای تعمیم اندازه گیری کنید. تعمیم به ظرفیت مدل برای به کارگیری آنچه از داده های آموزشی آموخته است برای پیش بینی های دقیق بر روی نمونه های جدید و نادیده اشاره دارد – یک جنبه حیاتی از یادگیری ماشین.

به طور خلاصه، تقسیم یک مجموعه داده به زیرمجموعه های آموزشی و آزمایشی یک گام مهم در گردش کار یادگیری ماشین، از جمله NLP است. این به شما امکان می دهد عملکرد مدل را ارزیابی کنید، بیش از حد برازش را تشخیص دهید، فراپارامترها را تنظیم کنید، بهترین مدل را انتخاب کنید، و اطمینان حاصل کنید که مدل شما می تواند به داده های جدید و نادیده تعمیم یابد – جنبه های اساسی ساخت مدل های یادگیری ماشین موثر و قابل اعتماد.

در این پروژه ما با استفاده کد های زیر ابتدا ۴ لیبل از احساسات موجود در دیتاست را جدا میکنیم.

```
df = pd.read_csv('pre_proccessed.csv')
desiered_labels = ['neutral', 'sadness', 'worry', 'happiness']
filtered_df = df[df['emotions'].isin(desiered_labels)]
filtered_df.to_csv('4_emotions_data.csv')
```

در ادامه label ها را encode کرده

```
# Load data frame
df = pd.read_csv('4_emotions_data.csv')
df.dropna(inplace=True)
X = df['joined_lemmatized']
y = df['emotions']

# Encode the labels into numerical values
label_encoder = LabelEncoder()
encoded_labels = label_encoder.fit_transform(y)
```

و سپس داده‌ها را به دو بخش test, train تقسیم میکنیم

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, encoded_labels,
test_size=0.2, random_state=42)
```

مرحله بعدی vectorize کردن داده‌ها است که قطعه کد زیر آین امر را میسر میکند

```
# Create a TF-IDF vectorizer
vectorizer = TfidfVectorizer()
```

```
# Vectorize the text data
X_train_vectorized = vectorizer.fit_transform(X_train)
X_test_vectorized = vectorizer.transform(X_test)
```

بعد از طی کردن این مراحل تمام کار ما با پیش پردازش داده‌ها به اتمام رسیده و دو مجموعه x_train, y_train آماده تزریق به الگوریتم هایی هستند که کار یادگیری را انجام میدهند.

۳.۴ ساختن مدل

در زمینه یادگیری ماشینی، مدل یک نمایش ریاضی یا محاسباتی است که الگوها و روابط را از داده ها یاد می گیرد. این به عنوان یک سیستم الگوریتمی عمل می کند که می تواند پیش بینی ها، طبقه بندی ها یا دیگر استنباط ها را بر روی داده های جدید و نادیده بر اساس آنچه در طول آموزش آموخته است انجام دهد.

یک مدل اساساً تابعی است که داده های ورودی را به پیش بینی های خروجی نگاشت می کند. به عنوان مثال، در NLP، یک مدل ممکن است یک قطعه از متن را بگیرد و احساسات، دسته‌بندی یا هر اطلاعات مرتبط دیگری را پیش‌بینی کند.

ایجاد یک مدل:

۱. **انتخاب یک معماری مدل**: اولین قدم در ایجاد یک مدل، انتخاب یک معماری مناسب است. در NLP، این می تواند مدل های مختلفی باشد، مانند:
۲. **مدل‌های کلاسیک** (مانند Naive Bayes، SVM): این مدل‌ها بر اساس اصول آماری و ریاضی سنتی هستند.
۳. **شبکه های عصبی** (به عنوان مثال، LSTM ،RNN، ترانسفورماتور): این مدل ها بر اساس شبکه های عصبی مصنوعی هستند که از ساختار مغز انسان الهام گرفته شده اند.
۴. **آماده سازی داده ها**: قبل از آموزش یک مدل، داده ها باید آماده شوند. این شامل کارهایی مانند تمیز کردن متن، توکن سازی، برداری (تبدیل متن به فرمت عددی) و تقسیم داده ها به مجموعه های آموزشی و آزمایشی است.

۵. **مهندسی ویژگی (اختیاری)**: در NLP، مهندسی ویژگی شامل تبدیل داده‌های متن خام به قالبی است که برای مدل انتخابی مناسب است. این ممکن است شامل تکنیک هایی مانند جاسازی کلمه، TF-IDF یا سایر نمایش ها باشد.

۶. **کامپایل و آموزش مدل**: برای مدل‌های کلاسیک، این مرحله شامل تنظیم مدل، تعیین فراپارامترها و سپس برازش آن با داده‌های آموزشی است. برای شبکه های عصبی، این شامل تعریف معماری، کامپایل مدل و آموزش آن با استفاده از یک الگوریتم بهینه سازی است.

۷. **ارزیابی مدل**: هنگامی که مدل آموزش داده شد، برای ارزیابی عملکرد آن باید در مجموعه تست ارزیابی شود. معیارهای ارزیابی رایج در NLP شامل دقت، دقت، یادآوری، امتیاز F۱ و معیارهای خاص‌تر بسته به کار است (به عنوان مثال، امتیاز BLEU برای ترجمه ماشینی).

۸. **تنظیم و بهینه سازی**: بسته به نتایج ارزیابی، ممکن است نیاز به تنظیم دقیق مدل داشته باشید. این می تواند شامل تنظیم هایپرپارامترها، تغییر معماری، یا انجام تکنیک هایی مانند منظم سازی برای بهبود عملکرد باشد.

۹. **استقرار و استنتاج**: هنگامی که مدل به طور رضایت بخشی عمل کرد، می توان آن را برای استفاده در دنیای واقعی مستقر کرد. این شامل دریافت داده‌های جدید و پیش‌بینی یا طبقه‌بندی بر اساس آموخته‌های آن است.

۱۰. **نظارت و نگهداری**: پس از استقرار، نظارت بر عملکرد مدل در طول زمان بسیار مهم است. اگر توزیع داده های اساسی تغییر کند یا اگر عملکرد مدل بدتر شود، ممکن است نیاز به بازآموزی یا تنظیم دقیق باشد.

ما در این پروژه از مدل های زیر استفاده میکنیم:

۱. SVC
۲. Random Forest
۳. XGBoost
۴. Naive Bayes

در اداه به توضیح این مدل ها میپردازیم.

SVC ۳.۵.۱

SVC مخفف عبارت Support Vector Classification است. این یک الگوریتم یادگیری ماشینی نظارت شده است که برای کارهای طبقه بندی استفاده می شود. طبقه بندی بـردار پشـتیبان بـرای طبقه بندی به عنوان SVM (ماشین بردار پشتیبانی) نیز شناخته می شود.

۱. **وظیفه طبقه بندی**: هدف اصلی مدل طبقه بندی بردار پشتیبانی (SVC) طبقه بنـدی نقـاط داده به یکی از دو یا چند کلاس است. این به ویژه برای مسائل طبقه بندی باینری مفید است، جـایی که هدف تخصیص نقاط داده به یکی از دو کلاس است، اما می توان آن را به طبقه بنـدی چنـد کلاسه گسترش داد.
۲. **حداکثر کردن حاشیه**: SVC با یافتن یک مرز تصمیم (یا ابرصفحه) کار می کند که نقـاط داده کلاس های مختلف را به بهترین شکل جدا می کند و در عین حال حاشیه بین کلاس هـا را بـه حداکثر می رساند. حاشیه فاصله بین ابر صفحه و نزدیکترین نقاط داده هر کلاس است. حداکثر کردن حاشیه به بهبود تعمیم مدل کمک می کند و خطر بیش از حد برازش را کاهش می دهد.
۳. **بردارهای پشتیبان**: بردارهای پشتیبان، نقاط داده ای هستند که بـه مـرز تصـمیم نزدیـک تـرین هستند. اینها مهمترین نقاط در مجموعه داده هستند زیرا آنهایی هستند کـه بـر مـوقعیت و جهت گیری مرز تصمیم تأثیر می گذارند. الگوریتم "بردار پشتیبانی" نامگذاری شده است زیرا بـر این نقاط داده کلیدی متکی است.
۴. **ترفند هسته**: SVC می تواند با استفاده از تکنیکی به نام "ترفند هسته"، هم مجموعه داده های قابل جداسازی خطی و هم مجموعه داده های غـیرخطی را مـدیریت کنـد. این تکنیـک فضـای ویژگی اصلی را به فضایی با ابعاد بالاتر تبدیل می کند که در آن نقاط داده به صورت خطی قابل تفکیک می شوند. توابع رایج هسته شامل هسته خطی، هسته چند جملـه ای و هسـته تـابع پایـه شعاعی (RBF) است.
۵. **پارامتر C**: الگوریتم SVC دارای یک فراپارامتر به نام "C" (پارامتر منظم سـازی) اسـت کـه مبادله بین حداکثر کردن حاشیه و به حداقل رساندن خطای طبقه بندی را کنترل می کنـد. مقـدار کمتر C اجازه می دهد تا حاشیه وسیع تری داشته باشد، اما ممکن است منجـر بـه طبقـه بنـدی

اشتباه برخی از نقاط داده شود، در حالی که مقدار بزرگتر C منجر به حاشیه باریک تر اما طبقه بندی اشتباه کمتر می شود.

۶. **Hyperplane**: در یک مسئله طبقه بندی باینری (دو کلاس)، مدل SVC یک ابر صفحه را پیدا می کند که داده ها را به دو کلاس جدا می کند. برای مسائل چند کلاسه، از چند ابرصفحه برای جداسازی جفت‌های مختلف کلاس استفاده می‌شود.

۷. **پیش بینی**: زمانی که مدل SVC بر روی یک مجموعه داده برچسب دار آموزش داده شد، می توان از آن برای پیش بینی نقاط داده جدید و بدون برچسب استفاده کرد. این مدل نقاط داده جدید را به یکی از کلاس ها بر اساس موقعیت آنها نسبت به مرز تصمیم اختصاص می دهد.

مزایای کلیدی SVC:

- SVC در فضاهای با ابعاد بالا موثر است.
- می تواند مشکلات طبقه بندی خطی و غیر خطی را حل کند.
- استفاده از بردارهای پشتیبانی به تمرکز بر آموزنده ترین نقاط داده کمک می کند.
- هنگامی که تنظیم مناسب اعمال می شود در برابر بیش از حد برازش قوی است.

محدودیت های SVC:

- SVC می تواند از نظر محاسباتی گران باشد، به خصوص با مجموعه داده های بزرگ.
- انتخاب هسته و فراپارامترها می تواند بر عملکرد مدل تأثیر بگذارد و نیاز به تنظیم دقیق دارد.
- ممکن است در مجموعه داده هایی با کلاس های نویزدار یا همپوشانی خوب عمل نکند.

به طور خلاصه، دسته‌بندی بردار پشتیبانی (SVC) یک الگوریتم یادگیری ماشینی قدرتمند است که برای کارهای طبقه‌بندی استفاده می‌شود، به‌ویژه زمانی که هدف یافتن مرز تصمیم‌گیری است که حاشیه بین کلاس‌ها را به حداکثر می‌رساند. این می تواند مشکلات طبقه بندی خطی و غیرخطی را از طریق استفاده از توابع هسته حل کند و زمانی موثر است که داده ها ابعاد بالایی داشته باشند. با این حال، به تنظیم دقیق پارامتر نیاز دارد و می تواند محاسباتی فشرده باشد.

در این بخش ما با استفاده از قطعه کد زیر یک مدل **SVC** میسازیم و داده‌ها را به آن تزریق میکنیم

```
# SVC model
# Create SVC model
svc_model = SVC()

# Train the model
svc_model.fit(X_train_vectorized, y_train)

# Make predictions on the test set
y_pred_svc = svc_model.predict(X_test_vectorized)

# Decode the predicted labels
decoded_pred = label_encoder.inverse_transform(y_pred_svc)

# Save the model to a file
filename_svc = 'svc_model.sav'
joblib.dump(svc_model, filename_svc)
```

در این قطعه کد ما ابتدا مدل یک شی از جنس مدل **SVC** میسازیم.

داده‌های مربوط به یادگیری را به مدل پاس میدهیم.

دقت یادگیری مدل را با استفاده از داده‌های تست اندازه‌گیری میکنیم.

لیبل های داده‌های تست را دیکد میکنیم که بعداً بتوانیم برای بررسی آن‌ها را تشخیص دهیم.

و در نهایت هم مدل را ذخیره میکنیم.

۳.۵.۲ Random Forest

مدل جنگل تصادفی یک روش یادگیری گروهی است که هم برای کارهای طبقه‌بندی و هم رگرسیون در یادگیری ماشین استفاده می‌شود. این یک الگوریتم همه کاره و قدرتمند است که به دلیل توانایی خود در ارائه پیش بینی های دقیق و قوی محبوبیت پیدا کرده است.

مفاهیم کلیدی:

۱. **یادگیری گروهی**: جنگل تصادفی یک تکنیک یادگیری گروهی است، به این معنی که پیش‌بینی‌های چند مدل پایه (درخت تصمیم) را برای بهبود عملکرد کلی ترکیب می‌کند. به جای تکیه بر یک درخت تصمیم، Random Forest از خرد جمعیت برای پیش‌بینی دقیق‌تر استفاده می‌کند.

۲. **درختان تصمیم**: در هسته یک جنگل تصادفی، درختان تصمیم فردی قرار دارند. درخت های تصمیم سازه های ساده و درخت مانندی هستند که بر اساس مجموعه ای از قوانین آموخته شده از داده ها تصمیم می گیرند. هر درخت تصمیم در Random Forest به طور مستقل ساخته می شود و مجموعه قوانین خاص خود را دارد.

۳. **تصادفی سازی**: "تصادفی" در جنگل تصادفی به دو منبع کلیدی تصادفی مورد استفاده در ساخت آن اشاره دارد:

۴. **نمونه‌گیری تصادفی**: هر درخت تصمیم بر روی زیرمجموعه‌ای تصادفی از داده‌های آموزشی با جایگزینی (نمونه‌گیری بوت استرپ) آموزش داده می‌شود. این باعث ایجاد تنوع در بین درختان می شود.

۵. **انتخاب ویژگی تصادفی**: در هر گره درخت تصمیم، یک زیرمجموعه تصادفی از ویژگی ها هنگام تصمیم گیری تقسیم در نظر گرفته می شود. این باعث تنوع بیشتر و کاهش بیش از حد برازش می شود.

۶. **رای‌گیری یا میانگین‌گیری**: برای کارهای طبقه‌بندی، Random Forest پیش‌بینی‌های درختان منفرد را با رای اکثریت ترکیب می‌کند. به عبارت دیگر، کلاسی را انتخاب می کند که بیشترین رای را از درختان جداگانه دریافت کند. برای وظایف رگرسیون، میانگین مقادیر پیش‌بینی‌شده را از همه درخت‌ها می‌گیرد.

مزایای جنگل تصادفی:

- **دقت بالا**: Random Forest به دلیل دقت پیش‌بینی بالا معروف است. با جمع‌آوری پیش‌بینی‌ها از چندین درخت، خطر بیش از حد برازش را کاهش می‌دهد و نتایج قوی ارائه می‌کند.
- **انتخاب ویژگی ضمنی**: فرآیند انتخاب ویژگی تصادفی می تواند به عنوان مکانیزم انتخاب ویژگی ضمنی عمل کند. ویژگی هایی که به طور مداوم در بالای درختان تصمیم ظاهر می شوند، احتمالاً در پیش بینی ها مهم هستند.
- **مقاومت در برابر برازش بیش از حد**: ترکیبی از تکنیک‌های تصادفی‌سازی و میانگین‌گیری در میان درختان، جنگل تصادفی را حتی در مجموعه داده‌های پر سر و صدا یا با ابعاد بالا در برابر بیش‌برازش مقاوم می‌کند.
- **مدیریت مقادیر از دست رفته**: Random Forest می تواند مقادیر از دست رفته در مجموعه داده را مدیریت کند. این نیازی به نسبت داده های از دست رفته ندارد، و همچنان می تواند پیش بینی های معناداری ارائه دهد.
- **اهمیت ویژگی**: می‌تواند اهمیت هر یک از ویژگی‌ها را در مجموعه داده تخمین بزند و به تشخیص اینکه کدام متغیرها بیشترین سهم را در پیش‌بینی‌ها دارند کمک می‌کند.

محدودیت های جنگل تصادفی:

- **مدل جعبه سیاه**: جنگل تصادفی را می توان یک مدل جعبه سیاه در نظر گرفت، و تفسیر منطق تصمیم خاص مورد استفاده توسط درختان را به چالش می کشد.
- **محاسباتی فشرده**: آموزش یک جنگل تصادفی بزرگ با درختان زیاد می تواند از نظر محاسباتی گران باشد و ممکن است به حافظه قابل توجهی نیاز داشته باشد.
- **برای همه داده ها مناسب نیست**: در حالی که Random Forest یک الگوریتم همه کاره است، ممکن است بهترین انتخاب برای همه انواع داده ها و مشکلات نباشد.

به طور خلاصه، مدل جنگل تصادفی یک تکنیک یادگیری گروهی قدرتمند است که قدرت درخت‌های تصمیم‌گیری متعدد را برای ارائه پیش‌بینی‌های دقیق و قوی ترکیب می‌کند. این به طور گسترده در برنامه های مختلف یادگیری ماشین، از جمله وظایف طبقه بندی و رگرسیون، که در آن عملکرد بالا و انعطاف پذیری در برابر بیش از حد مطلوب است، استفاده می شود.

قطعه کد استفاده شده برای ساختن این مدل نیز همانند **SVC** است و تفاوتی ندارد که در زیر آمده است.

```
# Random Forest model
# Create a Random Forest model
rf_model = RandomForestClassifier()

# Train the model
rf_model.fit(X_train_vectorized, y_train)

# Make predictions on the test set
y_pred_rf = rf_model.predict(X_test_vectorized)

# Decode the predicted labels
decoded_pred = label_encoder.inverse_transform(y_pred_rf)

# Save the model to a file
filename_rf = 'random_forest_model.sav'
joblib.dump(rf_model, filename_rf)
```

XGBoost ۳.۵.۳

XGBoost، مخفف Extreme Gradient Boosting، یک الگوریتم یادگیری ماشینی است که به خانواده روش‌های تقویت گرادیان تعلق دارد. این به دلیل عملکرد پیش بینی استثنایی خود مشهور است و به طور گسترده در مسابقات مختلف یادگیری ماشین و برنامه های کاربردی در دنیای واقعی استفاده می شود.

مفاهیم کلیدی:

۱. تقویت گرادیان: XGBoost یک تکنیک یادگیری گروهی است که با ترکیب پیش‌بینی‌های چند زبان‌آموز ضعیف، معمولاً درخت‌های تصمیم، یک مدل پیش‌بینی قوی ایجاد می‌کند. این یک رویکرد تقویتی را دنبال می کند، که در آن هر یادگیرنده جدید بر روی اشتباهات انجام شده توسط افراد قبلی تمرکز می کند.

۲. **Decision Trees به عنوان Base Learners**: در XGBoost از درخت های تصمیم به عنوان پایه یا زبان آموزان ضعیف استفاده می شود. این درختان کم عمق با تعداد گره و عمق محدود هستند. ایده این است که مجموعه ای از چنین درختانی بسازیم که با هم کار کنند تا پیش بینی های دقیقی انجام دهند.

۳. **بهینه‌سازی گرادیان نزولی**: XGBoost از بهینه‌سازی گرادیان نزول برای به حداقل رساندن عملکرد تلفات استفاده می‌کند. تابع ضرر تفاوت بین مقادیر پیش بینی شده و واقعی را اندازه گیری می کند. با تنظیم مکرر پارامترهای درختان، XGBoost این تابع از دست دادن را به حداقل می رساند و منجر به پیش بینی های بهتر می شود.

۴. **منظم سازی**: برای جلوگیری از برازش بیش از حد، XGBoost از تکنیک های منظم سازی، از جمله منظم سازی (Lasso) L۱ و (Ridge) L۲ استفاده می کند. این جریمه ها برای پیچیدگی درختان و وزن های اختصاص داده شده به ویژگی ها اعمال می شود.

۵. **اهمیت ویژگی**: XGBoost معیاری از اهمیت ویژگی ارائه می دهد که به شناسایی تأثیرگذارترین ویژگی ها در پیش بینی ها کمک می کند. این اطلاعات می تواند انتخاب ویژگی و مهندسی را راهنمایی کند.

۶. **Cross-Validation**: اعتبار متقاطع اغلب با XGBoost برای ارزیابی عملکـــرد آن و تنظیم هایپرپارامترها استفاده می شود. این شامل تقسیم داده ها به زیر مجموعه هـــای متعـدد و آموزش مدل بر روی ترکیب های مختلف مجموعه های آموزشی و اعتبار سنجی است.

مزایای XGBoost:

- **دقت پیش‌بینی بالا**: XGBoost برای دقت پیش‌بینی پیشرفته‌اش شناخته شده است. اغلب در طیف وسیعی از کارها از سایر الگوریتم های یادگیری ماشین بهتر عمل می کند.
- **کارایی**: XGBoost برای کارایی طراحی شده است و می تواند مجموعه داده هـای بـزرگ و فضاهای ویژگی با ابعاد بالا را مدیریت کند.
- **استحکام**: به لطف تکنیک های منظم سازی که آن را برای مجموعه داده های پر سر و صـــدا یـــا پیچیده مناسب می کند، نسبت به بیش از حد برازش قوی است.
- **انعطاف‌پذیری**: XGBoost را می‌توان برای کارهای طبقه‌بندی و رگرسیون استفاده کـرد و آن را برای مشکلات مختلف یادگیری ماشین همه‌کاره می‌سازد.
- **اهمیت ویژگی**: این مدل بینش هایی در مورد اهمیت ویژگی ارائه می دهـد و بـــه دانشـــمندان و تحلیلگران داده کمک می کند تا تأثیرگذارترین متغیرها را در داده ها درک کنند.

محدودیت های XGBoost:

- **پیچیدگی**: XGBoost می‌تواند فراپارامترهای زیادی برای تنظیم داشته باشد که بهینه‌سازی آن را نسبت به الگوریتم‌های ساده‌تر پیچیده‌تر می‌کند.
- **منابع محاسباتی**: آموزش مجموعه های بزرگ درختـان در XGBoost می توانـد محاسـباتی فشرده باشد و ممکن است به منابع قابل توجهی نیاز داشته باشد.
- **مدل جعبه سیاه**: مانند سایر روش‌های مجموعه، XGBoost را می‌توان یک مدل جعبه ســـیاه در نظر گرفت، که تفسیر منطق تصمیم‌گیری خاص درختان را به چالش می‌کشد.

به طور خلاصه، XGBoost یک الگوریتم یادگیری ماشینی بسیار موثر و همه کاره است که در دقت پیش‌بینی برتر است. این به طور گسترده در برنامه‌های مختلف علوم داده و یادگیری ماشین، از جمله مسابقات Kaggle استفاده می‌شود و برای ارائه عملکرد بالاتر در طیف وسیعی از وظایف، شهرت دارد.

کد استفاده شده به شرح زیر است که مشابه با مدل های قبلیست.

```
# XGBoost model
# Create an XGBoost model
xgb_model = xgb.XGBClassifier()

# Train the model
xgb_model.fit(X_train_vectorized, y_train)

# Make predictions on the test set
y_pred_xgb = xgb_model.predict(X_test_vectorized)

# Calculate the accuracy
accuracy_xgb = accuracy_score(y_test, y_pred_xgb)

# Save the model to a file
filename_xgb = 'gradient_boosting_model.sav'
joblib.dump(xgb_model, filename_xgb)
```

۳.۵.۴ Naive Bayes

Naive Bayes یک الگوریتم یادگیری ماشین احتمالی است که معمولاً برای کارهای طبقه بندی استفاده می شود. این بر اساس قضیه بیز است و یک فرض کلیدی برای استقلال ویژگی ایجاد می کند، به همین دلیل است که به آن "ساده لوح" می گویند. با وجود سادگی، Naive Bayes می تواند به طرز شگفت انگیزی در پردازش زبان طبیعی (NLP) و وظایف طبقه بندی متن موثر باشد.

مفاهیم کلیدی:

۱. **قضیه بیز**: در هسته الگوریتم ساده بیز قضیه بیز قرار دارد که یک قضیه اساسی در نظریه احتمالات است. احتمال وقوع یک رویداد را بر اساس دانش قبلی از شرایطی که ممکن است با رویداد مرتبط باشد، توصیف می کند. در طبقه بندی، به محاسبه احتمال یک کلاس با توجه به مجموعه ای از ویژگی ها کمک می کند.
۲. **استقلال مشروط**: فرض "ساده لوح" در Naive Bayes این است که ویژگی ها با توجه به برچسب کلاس مستقل هستند. این بدان معنی است که وجود یک ویژگی مستقل از وجود هر ویژگی دیگر فرض می شود. اگرچه این فرض ممکن است در واقعیت صادق نباشد، اما Naive Bayes هنوز هم می تواند در عمل عملکرد خوبی داشته باشد.
۳. **چند جمله ای و گاوسی ساده بیز**: انواع مختلفی از مدل های نایو بیز وجود دارد.
   - Multinomial Naive Bayes: این نوع معمولاً برای کارهای طبقه‌بندی متن استفاده می‌شود که در آن ویژگی‌ها تعداد کلمات یا فرکانس‌های عبارت را نشان می‌دهند. این برای داده های گسسته مانند تعداد کلمات در یک سند مناسب است.
   - Gaussian Naive Bayes: این نوع فرض می کند که ویژگی ها از توزیع گاوسی (عادی) پیروی می کنند. برای ویژگی های با ارزش پیوسته استفاده می شود.
۴. **تخمین احتمال ویژگی**: Naive Bayes توزیع احتمال ویژگی ها را برای هر کلاس در داده های آموزشی تخمین می زند. به عنوان مثال، در طبقه بندی متن، احتمال وقوع هر کلمه در اسناد یک کلاس خاص را محاسبه می کند.

۵. Naive Bayes :Class Prior Probability همچنین احتمال قبلی هر کلاس را تخمین می زند. این احتمال تعلق یک سند به یک کلاس خاص بر اساس توزیع کلی کلاس ها در داده های آموزشی است.

مزایای Naive Bayes:

- **کارایی**: Naive Bayes از نظر محاسباتی کارآمد است و با مجموعه داده های بزرگ و فضاهای ویژگی با ابعاد بالا به خوبی مقیاس می شود.
- **آموزش سریع و پیش‌بینی**: سادگی الگوریتم منجر به آموزش سریع و زمان‌های پیش‌بینی می‌شود و آن را برای برنامه‌های بلادرنگ یا نزدیک به زمان مناسب می‌سازد.
- **عملکرد خوب در طبقه بندی متن**: Naive Bayes اغلب در کارهای طبقه بندی متن مانند تشخیص هرزنامه، تجزیه و تحلیل احساسات و دسته بندی اسناد استفاده می شود. این می تواند واژگان بزرگ و داده های متنی با ابعاد بالا را به طور موثر اداره کند.
- **نیاز به داده محدود:** حتی با مقدار کمی از داده های آموزشی می تواند عملکرد مناسبی داشته باشد.

محدودیت های Naive Bayes:

- **فرض مستقل بودن ویژگی**: فرض "ساده لوحانه" استقلال ویژگی ممکن است در همه موارد صادق نباشد. در عمل، اغلب همبستگی بین ویژگی ها وجود دارد
- **حساسیت به کیفیت داده ها**: Naive Bayes می تواند به کیفیت داده ها حساس باشد. ممکن است با ویژگی‌های نادر یا دیده نشده در داده‌های آزمایشی مشکل داشته باشد.
- **فقدان تفسیرپذیری مدل**: مانند سایر مدل‌های احتمالی، Naive Bayes یک مدل جعبه سیاه در نظر گرفته می‌شود، به این معنی که ممکن است بینش قابل تفسیری در مورد روابط بین ویژگی‌ها ارائه نکند.

به طور خلاصه، Naive Bayes یک الگوریتم ساده و در عین حال قدرتمند است که به ویژه برای کارهای طبقه بندی متن مناسب است. کارایی، سرعت و توانایی آن در مدیریت داده های با ابعاد بالا، آن

را به یک انتخاب محبوب تبدیل کرده است، به ویژه در برنامه های NLP مانند فیلتر کردن هرزنامه، تجزیه و تحلیل احساسات و دسته بندی اسناد.

برای این مدل هم مطابق مدل های قبلی عمل میکنیم.

```
# Naive Bayes model
# Create a Naive Bayes model
nb_model = MultinomialNB()

# Train the model
nb_model.fit(X_train_vectorized, y_train)

# Make predictions on the test set
y_pred_nb = nb_model.predict(X_test_vectorized)

# Calculate the accuracy
report_nb = classification_report(y_test, y_pred_nb)

# Save the model to a file
filename_nb = 'naive_bayes_model.sav'
joblib.dump(nb_model, filename_nb)
```

۴ ارزیابی روش

در بخش‌های قبلی ما در قطعه کد های خود از معیار دقت استفاده کردیم که قطعه کد آن بـه شـرح زیـر است:

```
accuracy_xgb = accuracy_score(y_test, y_pred_xgb)
```

اما علاوه بر معیار دقت که تنها به بررسی تعداد حدس های درسـت بـر غلـط اسـتوار اسـت معیارهـای دیگری نیز برای بررسی وجود دارد

در ادامه به بررسی چندی از روش‌های ارزیابی میپردازیم:

۴.۱ accuracy

دقت پیش بینی های مثبت را اندازه گیری می کند. این نسبت پیش‌بینی‌هـای مثبت واقعی بـه تعـداد کـل پیش‌بینی‌های مثبت (مثبت واقعی + مثبت کاذب) است. دقت زمانی ارزشمند است که هزینه مثبت کاذب بالا باشد.

```
accracy = TP / (TP + FP)
```

۴.۲ Recall

یادآوری توانایی مدل را در شناسایی صحیح تمام موارد مثبت اندازه گیری می کند. این نسبت پیش بیـنی های مثبت واقعی به تعداد کل نمونه های مثبت واقعی (مثبت‌هـای واقعی + منفی هـای کـاذب) اسـت. یادآوری زمانی ارزشمند است که از دست دادن موارد مثبت پرهزینه باشد.

```
recall = TP / (TP + FN)
```

F1-Score ۴.۳

F۱-Score میانگین هارمونیک دقت و یادآوری است. این یک معیار متعادل از عملکرد یک مدل را ارائه می دهد، به خصوص زمانی که عدم تعادل بین کلاس ها وجود دارد. به صورت زیر محاسبه می شود:

```
F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
```

Specificity ۴.۴

Specificity توانایی مدل را در شناسایی صحیح موارد منفی می سنجد. این نسبت پیش‌بینی‌های منفی واقعی به تعداد کل نمونه‌های منفی واقعی (منفی واقعی + مثبت کاذب) است. زمانی که اجتناب از هشدارهای اشتباه برای موارد منفی بسیار مهم است، ویژگی ارزشمند است.

```
Specificity = TN / (TN + FP)
```

False Positive Rate (FPR) ۴.۵

نسبت پیش بینی های مثبت کاذب به تعداد کل موارد منفی واقعی است. مکمل ویژگی است و به صورت زیر محاسبه می شود:

```
FPR = FP / (TN + FP)
```

False Negative Rate (FNR) ۴.۶

نرخ منفی کاذب (FNR): FNR نسبت پیش بینی های منفی کاذب به تعداد کل نمونه های مثبت واقعی است. نرخی را که مدل در آن موارد مثبت را از دست می دهد اندازه گیری می کند و به صورت زیر محاسبه می شود:

```
FNR = FN / (TP + FN)
```

۴.۷ Area Under the Receiver Operating Characteristic (ROC AUC)

اندازه گیری توانایی مدل در تشخیص نمونه های مثبت و منفی است. این ناحیه زیر منحنی مشخصه عملیاتی گیرنده را کمی می کند، با مقادیر بالاتر که نشان دهنده تبعیض بهتر است.

این معیارها نمای جامعی از عملکرد مدل شما ارائه می دهند، به خصوص در سناریوهایی با مجموعه داده های نامتعادل یا جایی که انواع مختلف خطاها هزینه های متفاوتی دارند.

برای محاسبه این معیارها در پایتون با استفاده از scikit-learn، می‌توانید از تابع classification_report استفاده کنید که خلاصه‌ای از دقت، یادآوری، امتیاز F۱ و سایر معیارها را برای هر کلاس در مسئله طبقه‌بندی شما ارائه می‌کند.

```
from sklearn.metrics import classification_report

# true_labels are the true class labels from the test set
# predicted_labels are the predicted class labels from your model
report = classification_report(true_labels, predicted_labels)
print(report)
```

به عنوان مثال خروجی به این شکل خواهد بود:

```
              precision    recall  f1-score   support

     Class 0       0.82      0.90      0.86       100
     Class 1       0.74      0.65      0.69        50
     Class 2       0.91      0.93      0.92        75

    accuracy                           0.84       225
   macro avg       0.82      0.83      0.82       225
weighted avg       0.84      0.84      0.84       225
```

در این مثال، ما یک مشکل طبقه بندی با سه کلاس داریم: کلاس ۰، کلاس ۱، و کلاس ۲. معنی هر قسمت از گزارش در اینجا آمده است:

- Precision: این ستون دقت هر کلاس را نشان می دهد. به عنوان مثال، برای کلاس ۰، دقت ۰.۸۲ است، به این معنی که ۸۲٪ از نمونه های پیش بینی شده به عنوان کلاس ۰ درست بوده اند.
- Recall: این ستون فراخوان (یا نرخ مثبت واقعی) را برای هر کلاس نشان می دهد. برای کلاس ۰، فراخوانی ۰.۹۰ است، که نشان می دهد ۹۰٪ از نمونه های کلاس ۰ واقعی به درستی توسط مدل شناسایی شده اند.
- F۱-Score: F۱-Score میانگین هارمونیک دقت و یادآوری است. تعادلی بین دقت و یادآوری ایجاد می کند. برای کلاس ۰، امتیاز ۰.۸۶ F۱ است.
- **پشتیبانی**: این ستون تعداد نمونه‌های هر کلاس را در مجموعه داده آزمایشی نشان می‌دهد.
- دقت: دقت در پایین گزارش نشان داده شده است. این دقت کلی مدل است که در این مورد ۰.۸۴ (۸۴٪) است.
- **میانگین ماکرو**: این ردیف میانگین کلان دقت، فراخوانی و امتیاز F۱ را ارائه می‌کند که در همه کلاس‌ها میانگین هستند. زمانی مفید است که بخواهید عملکرد مدل را در همه کلاس ها به طور یکسان ارزیابی کنید.
- **میانگین وزنی**: این ردیف با در نظر گرفتن عدم تعادل کلاس، دقت میانگین وزنی، فراخوانی و امتیاز F۱ را ارائه می دهد. زمانی مفید است که کلاس ها اندازه های متفاوتی داشته باشند و بخواهید وزن بیشتری به کلاس های بزرگتر بدهید.

classification_report خلاصه ای مختصر از عملکرد مدل شما برای هر کلاس و به طور کلی است. این یک ابزار ارزشمند برای درک اینکه مدل شما در کجا برتر است و در کجا ممکن است در یک مشکل طبقه بندی چند کلاسه نیاز به بهبود داشته باشد.

۵ نتایج

در این پایانامه و پروژه ما به بررسی مدل های مختلفی که برای پردازش زبان طبیعی وجود دارند پرداختیم و انواع الگوریتم هایی که میتوان به کمک آن‌ها مدلی بسازیم که نتایج را برای ما پیش‌بینی کنند بررسی کردیم. همچنین تعدادی از آن‌ها همراه با معایب و مزایای هرکدام مورد بحث و بررسی قرار گرفتند.

در انتها میتوان گفت مدل های یادگیری ماشین کارایی مناسبی برای پردازش زبان طبیعی دارند و تا زمانی که تعداد دسته بندی های ما زیاد نباشد عمل‌کرد خوبی به جا میگذارند اما با افزایش لیبل ها استفاده از الگوریتم های عصبی و عمیق خروجی بهتری برای ما خواهند داشت.

از طرفی خروجی ما وابسته به تعداد کلمات موجود در هر توییت است که با بیشتر شدن تعداد کلمات و vectorize بهتر میتوان خروجی بهتری گرفت.

در انتها میتوان گفت پارامتر های زیادی در خورجی ما تأثیر گذارند اما در صورتی که ما بتوانیم هر کلمه را به عددی بامعنا مپ کنیم و تبدیل کردن کلمات به اعداد بار احساسی آن‌ها را نیز منتقل کند خروجی ما به مراتب بهتر میشود.

همچنین، تکنیک‌های پیش پردازش و استخراج ویژگی تأثیر قابل‌توجهی بر عملکرد رویکردهای مختلف تحلیل احساسات و عواطف دارند.

۶ جمع‌بندی و کارهای آینده

کارهایی که در زمینه مپ کردن کلمات به اعداد بامعنا صورت بگیرند به خروجی بهتری منجر می‌شوند پس در آینده به روش‌های مختلف برای پیش پردازش داده‌ها بیشتر پرداخته خواهد شد. همچنین الگوریتم های مبتنی بر شبکه‌های عصبی و شبکه‌های عمیق هم مورد بحث و بررسی قرار خواهند گرفت. از سوی دیگر استفاده از سخت‌افزار های متفاوت برای بهتر شدن زمان یادگیری و ساختن مدل هم میتواند بسیاری از مشکلات ما را برای پیاده‌سازی الگوریتم های پیچیده از نظر زمانی حل کند.

7 مراجع

[1] Al Amrani Y, Lazaar M, El Kadiri KE (2018) Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Comput Sci 127:511–520

[2] Alqaryouti O, Siyam N, Monem AA, Shaalan K (2020) Aspect-based sentiment analysis using smart government review data. Appl Comput Inf. https://doi.org/10.1016/j.aci.2019.11.003

[3] Alswaidan N, Menai MEB (2020) A survey of state-of-the-art approaches for emotion recognition in text. Knowl Inf Syst 62(8):1–51

[4] Archana Rao PN, Baglodi K (2017) Role of sentiment analysis in education sector in the era of big data: a survey. Int J Latest Trends Eng Technol 22–24

[5] Arora M, Kansal V (2019) Character level embedding with deep convolutional neural network for text normalization of unstruc-tured data for twitter sentiment analysis. Soc Netw Anal Min 9(1):1–14

[6] Arulmurugan R, Sabarmathi K, Anandakumar H (2019) Classification of sentence level sentiment analysis using cloud machine learn-ing techniques. Cluster Comput 22(1):1199–1209

[7] Asghar MZ, Subhan F, Imran M, Kundi FM, Shamshirband S, Mosavi A, Csiba P, Várkonyi-Kóczy AR (2019) Performance evaluation of supervised machine learning techniques for efficient detection of emotions from online content. arXiv preprint arXiv:19080 1587

[8] Bakker I, Van Der Voordt T, Vink P, De Boon J (2014) Pleasure, arousal, dominance: Mehrabian and Russell revisited. Curr Psy-chol 33(3):405–421

[9] Balahur A, Turchi M (2014) Comparative experiments using super-vised learning and machine translation for multilingual sentiment analysis. Comput Speech Lang 28(1):56–75

[10] Bandhakavi A, Wiratunga N, Padmanabhan D, Massie S (2017) Lexi-con based feature extraction for emotion text classification. Pat-tern Recogn Lett 93:133–142

[11] Basiri ME, Abdar M, Cifci MA, Nemati S, Acharya UR (2020) A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques. Knowl Based Syst 198:105949

[12] Batbaatar E, Li M, Ryu KH (2019) Semantic-emotion neural network for emotion recognition from text. IEEE Access 7:111866–111878

[13] Becker K, Moreira VP, dos Santos AG (2017) Multilingual emotion classification using supervised learning: comparative experi-ments. Inf Process Manag 53(3):684–704

[14] Bernabé-Moreno J, Tejeda-Lorente A, Herce-Zelaya J, Porcel C, Her-rera-Viedma E (2020) A context-aware embeddings supported method to extract a fuzzy sentiment polarity dictionary. Knowl-Based Syst 190:105236

[15] Bhardwaj A, Narayan Y, Dutta M et al (2015) Sentiment analysis for Indian stock market prediction using sensex and nifty. Procedia Comput Sci 70:85–91

[16] Bhaskar J, Sruthi K, Nedungadi P (2015) Hybrid approach for emotion classification of audio conversation based on text and speech mining. Procedia Comput Sci 46:635–643

[17] Braun N, van der Lee C, Gatti L, Goudbeek M, Krahmer E (2021) Memofc: introducing the multilingual emotional football corpus. Lang Resour Eval 55(2):389–430

[18] Bučar J, Žnidaršič M, Povh J (2018) Annotated news corpora and a lexicon for sentiment analysis in Slovene. Lang Resour Eval 52(3):895–919

[19] Buechel S, Hahn U (2017) Emobank: Studying the impact of anno-tation perspective and representation format on dimensional emotion analysis. In: Proceedings of the 15th conference of the european chapter of the association for computational linguistics: volume 2, Short Papers, pp 578–585

[20] Chaffar S, Inkpen D (2011) Using a heterogeneous dataset for emotion analysis in text. In: Butz C, Lingras P (eds) Advances in artificial