



دانشکده مهندسی کامپیوتر

تشخیص مشکلات روحی از محتوای منتشر شده در شبکه‌های مجازی

پروژه برای دریافت درجه کارشناسی در رشته مهندسی کامپیوتر
گرایش هوش مصنوعی و رباتیک

آیلین نائب‌زاده

استاد راهنما

آقای دکتر سید صالح اعتمادی

شهریور ۱۴۰۳



تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پروژه

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: آیلین نائب‌زاده

عنوان پروژه: تشخیص مشکلات روحی از محتوای منتشر شده در شبکه‌های مجازی

تاریخ دفاع: شهریور ۱۴۰۳

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی و رباتیک

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
۱	استاد راهنما	دکتر صالح اعتمادی	استادیار	دانشگاه علم و صنعت ایران	
۲	استاد مدعو داخلی	دکتر ناصر مزینی	استادیار	دانشگاه علم و صنعت ایران	

تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب آیلین نائب‌زاده به شماره دانشجویی ۹۹۵۲۲۱۸۵ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی نتایج این پروژه حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. در صورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: آیلین نائب‌زاده

تاریخ و امضا:

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

☐ بهره‌برداری از این پایان‌نامه برای همگان بلامانع است.

☐ بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنما، بلامانع است.

☐ بهره‌برداری از این پایان‌نامه تا تاریخ ممنوع است.

استاد راهنما: آقای دکتر سید صالح

اعتمادی

تاریخ:

امضا:

تقدیم به:

پدر و مادرم

و

تمام اساتید گرامی و دوستانی که دانش خود را با
من به اشتراک گذاشتند.

قدردانی

سپاس خداوندگار حکیم را که با لطف بی‌کران خود، آدمی را زیور عقل آراست. در آغاز وظیفه خود می‌دانم از زحمات بی‌دریغ استاد راهنمای خود، جناب آقای دکتر اعتمادی، صمیمانه تشکر و قدردانی کنم که قطعاً بدون راهنمایی‌های ارزنده ایشان، این مجموعه به انجام نمی‌رسید. از جناب آقای صادق جعفری دانشجوی ارشد گرایش هوش مصنوعی که زحمت مطالعه و مشاوره این رساله را تقبل فرمودند و در آماده سازی این رساله، به نحو احسن اینجانب را مورد راهنمایی قرار دادند، کمال امتنان را دارم. در پایان، بوسه می‌زنم بر دستان خداوندگاران مهر و مهربانی، پدر و مادر عزیزم و بعد از خدا، ستایش می‌کنم وجود مقدس‌شان را و تشکر می‌کنم از خانواده عزیزم به پاس عاطفه سرشار و گرمای امیدبخش وجودشان، که بهترین پشتیبان من بودند.

آیلین نائب‌زاده

شهریور ۱۴۰۳

چکیده

این پایان‌نامه، به بحث در مورد تشخیص انواع بیماری‌های روحی و روان‌شناختی همانند افسردگی، اعتیاد، تمایل به خودکشی و استرس از محتوای منتشر شده توسط کاربران در شبکه‌های مجازی می‌پردازد. امروزه با رشد فناوری افراد بیشتری از شبکه‌های مجازی استفاده می‌کنند و روزانه تعداد بسیاری زیادی از مطالب گوناگون در انواع شبکه‌های اجتماعی منتشر می‌شود. همچنین یکی از دغدغه‌های افراد در صنایع گوناگون نرخ رو به افزایش افسردگی و سایر بیماری‌های روحی در میان انسان‌ها می‌باشد. در راستای این موضوع تشخیص سریع و درست این مشکلات دارای اهمیت فراوانی می‌باشد. حال در این پایان‌نامه قصد داریم تا الگوریتم‌های گوناگون حوزه یادگیری ماشین را بر روی تعدادی مجموعه داده جمع‌آوری شده اعمال کنیم و نتایج را با یک‌دیگر مقایسه کنیم. این مسئله یک وظیفه رده‌بندی متن می‌باشد که برای آن سعی شده است علاوه بر الگوریتم‌های قدیمی یادگیری ماشین همانند XGBoost و بردار ماشین پشتیبان (Support Vector Machine) از روش‌های به‌روزتر همانند مدل‌های زبانی بزرگ (Large Language Models) و مدل‌های (Transformers) خانواده BERT استفاده شود.

واژگان کلیدی: شبکه‌های مجازی، مشکلات روحی، رده‌بندی متن، یادگیری ماشین، XGBoost، بردار ماشین پشتیبان، مدل‌های زبانی بزرگ، مدل‌ها، BERT

فهرست مطالب

د	فهرست تصاویر	
ر	فهرست جداول	
۱	فصل ۱: مقدمه	
۱	۱-۱ مقدمه	
۳	فصل ۲: تعاریف و مفاهیم پایه	
۳	۲-۱ مقدمه	
۳	۲-۲ تعاریف و مفاهیم پایه	
۳	۲-۲-۱ هوش مصنوعی	
۴	۲-۲-۲ پردازش زبان طبیعی	
۴	۲-۲-۳ یادگیری نظارت شده	
۴	۲-۲-۴ استخراج متن	
۴	۲-۲-۵ پیش پردازش متن	
۵	۲-۲-۶ طبقه بندی متن	
۶	۲-۲-۷ ارزیابی مدل ها	
۸	۲-۲-۸ مدل	
۸	۲-۲-۹ بیش برآزش	
۸	۲-۲-۱۰ کم برآزش	
۹	۲-۲-۱۱ ارگرسیون لجستیک	

۱۰	۲-۲-۲ ماشین بردار پشتیبان
۱۰	۳-۲-۲ یادگیری گروهی
۱۰	۴-۲-۲ XGBoost
۱۱	۵-۲-۲ شبکه‌های عصبی مصنوعی
۱۳	۶-۲-۲ تابع فعال‌ساز
۱۴	۷-۲-۲ شبکه‌های عصبی عمیق
۱۴	۸-۲-۲ شبکه‌های عصبی پیچشی
۱۵	۹-۲-۲ شبکه‌های عصبی بازگشتی
۱۵	۱۰-۲-۲ شبکه‌های حافظه کوتاه‌مدت بلند
۱۶	۱۱-۲-۲ شبکه‌های حافظه کوتاه‌مدت بلند دو طرفه
۱۷	۱۲-۲-۲ مدل‌های زبانی بزرگ
۱۷	۱۳-۲-۲ مبدل‌ها
۱۸	۱۴-۲-۲ BERT
۱۹	۱۵-۲-۲ تنظیم دقیق
۲۰	فصل ۳: مروری بر کارهای مرتبط
۲۰	۱-۳ مقدمه
۲۰	۲-۳ کارهای پیشین
۲۰	۱-۲-۳ تشخیص و تحلیل با استفاده از الگوریتم‌های استخراج ویژگی در یادگیری ماشین
۲۴	۲-۲-۳ تشخیص و تحلیل با استفاده از مدل‌های زبانی بزرگ
۲۶	۳-۲-۳ چالش eRisk
۳۱	۴-۲-۳ چالش LT-EDI
۳۳	۵-۲-۳ نتیجه‌گیری
۳۴	فصل ۴: آزمایش‌های انجام شده
۳۴	۱-۴ مقدمه
۳۵	۲-۴ مجموعه داده Sentiment140

۳۵	۱-۲-۴ پیش‌پردازش
۳۷	۲-۲-۴ تجزیه داده‌ها به مجموعه آموزشی، تست و اعتبارسنجی
۳۷	۳-۲-۴ الگوریتم‌ها و تحلیل خروجی‌ها
۴۰	۳-۴ مجموعه داده DepressionTweet
۴۱	۱-۳-۴ پیش‌پردازش
۴۳	۲-۳-۴ تجزیه داده‌ها به مجموعه آموزشی، تست و اعتبارسنجی
۴۳	۳-۳-۴ الگوریتم‌ها و تحلیل خروجی‌ها
۴۷	۴-۴ مجموعه داده ویلاپرز و همکاران
۴۹	۱-۴-۴ پیش‌پردازش

۵۱	فصل ۵: نتیجه‌گیری و پیشنهادها
۵۱	۱-۵ نتیجه‌گیری و جمع‌بندی
۵۲	۲-۵ پیشنهادها
۵۲	۳-۵ کارهای آینده

۵۳	کتاب‌نامه
----	-----------

فهرست تصاویر

۱-۱	ساختار کلی پروژه چت بات	۲
۱-۲	اضلاع و متغیرهای ماتریس درهم ریختگی	۷
۲-۲	مثالی ساده از وقوع کم برآزش و بیش برآزش	۹
۳-۲	نمایی ساده از منحنی Logistic Regression	۹
۴-۲	نمایی ساده از نحوه کاربرد و معادله های SVM	۱۰
۵-۲	تعدادی از توابع فعال ساز در شبکه های عصبی به همراه روابط و نمودارها	۱۳
۶-۲	نمای کلی از شبکه های عصبی عمیق	۱۴
۷-۲	نمای کلی از شبکه های CNN	۱۵
۸-۲	نمای کلی از شبکه های عصبی بازگشتی	۱۶
۹-۲	نمای کلی از شبکه های BiLSTM	۱۷
۱۰-۲	معماری رمزگذار BERT	۱۸
۱-۳	مقدار امتیاز F1 محاسبه شده بر روی مجموعه داده توسط رده بندهای گوناگون در پژوهش [۱۲]	۲۱
۲-۳	معماری سیستم در پژوهش [۲۳]	۲۳
۳-۳	معماری شبکه CNN در پژوهش [۱۱]	۲۴
۴-۳	سیر زمانی و تاریخچه انتشار مبدل ها [۷]	۲۵
۵-۳	معماری سیستم در پژوهش [۱۰]	۲۶
۱-۴	نمایش ابر کلمات توییت ها	۳۶
۲-۴	پراکندگی تعداد کلمات در توییت های استخراج شده پس از پیش پردازش	۳۶

۳-۴	نمایش ماتریس‌های در هم‌ریختگی مدل‌های زبانی اعمال شده در آزمایش مجموعه داده
۳۹	Sent
۴-۴	نمایش نمودارهای ROC مدل‌های زبانی اعمال شده در آزمایش مجموعه داده Sent
۴۰	
۵-۴	توزیع داده‌ها براساس مقدار اولیه polarity در مجموعه داده DepressionTweet
۴۲	
۶-۴	توزیع داده‌ها براساس برچسب در مجموعه داده DepressionTweet
۴۲	
۷-۴	توزیع داده‌ها براساس برچسب در مجموعه داده DepressionTweet پس از اعمال داده‌افزایی
۴۳	
۸-۴	نمایش ماتریس‌های در هم‌ریختگی مدل‌های کلاسیک اعمال شده در آزمایش مجموعه داده
۴۵	DepressionTweet
۹-۴	پراکندگی تعداد کلمات در توییت‌های استخراج شده پس از پیش‌پردازش و داده‌افزایی
۴۶	
۱۰-۴	نمایش نمودارهای ROC مدل‌های زبانی اعمال شده در آزمایش مجموعه داده Depression
۴۷	Tweet
۱۱-۴	توزیع توییت‌های انگلیسی و اسپانیایی با توجه به برچسب‌ها پس از حذف برچسب‌های
۵۰	مشکل دار

فهرست جداول

۳۵	۱-۴ پنج ردیف اول مجموعه داده Sentiment140
۳۷	۲-۴ مدل‌های بر پایه استخراج ویژگی بر روی مجموعه داده Sent140 به همراه معیار صحت
۳۸	۳-۴ مقادیر برخی hyper parameter مدل‌های زبانی بر روی مجموعه داده Sent140
۳۸	۴-۴ مقایسه عملکرد LLMها با توجه به معیارهای صحت، دقت، یادآوری و امتیاز F1
۴۱	۵-۴ پنج ردیف اول مجموعه داده DepressionTweet
۴۱	۶-۴ پنج ردیف اول مجموعه داده DepressionTweet پس از پیش‌پردازش و اعمال تابع TextBlob
۴۴	۷-۴ مقایسه عملکرد مدل‌های کلاسیک با توجه به معیارهای صحت، دقت، یادآوری و امتیاز F1
۴۶	۸-۴ مقادیر برخی hyper parameter مدل‌های زبانی بر روی مجموعه داده Depression Tweet
۴۶	۹-۴ مقایسه عملکرد LLMها با توجه به معیارهای صحت، دقت، یادآوری و امتیاز F1
۴۸	۱۰-۴ پنج ردیف اول مجموعه داده ویلاپرز و همکاران

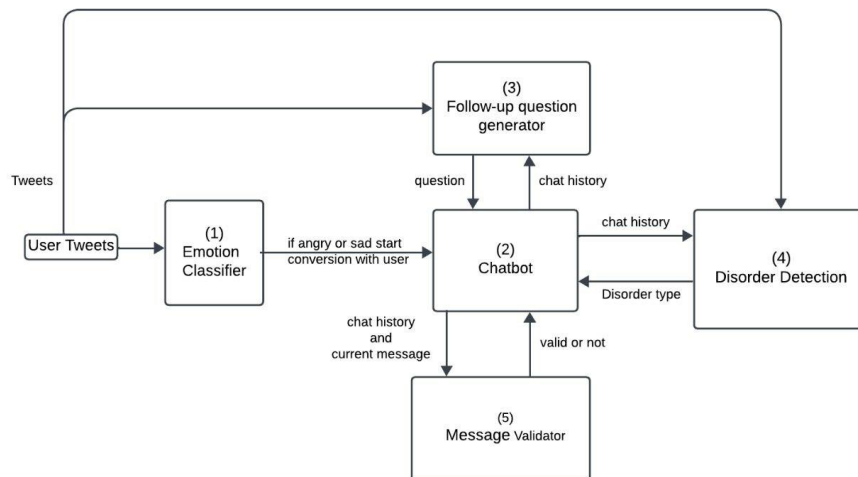
فصل ۱

مقدمه

۱-۱ مقدمه

امروزه با گسترش روزافزون استفاده از شبکه‌های اجتماعی، تعداد زیادی از افراد احساسات، مشکلات و تجربیات شخصی خود را به صورت آنلاین به اشتراک می‌گذارند. این پلتفرم‌ها نه تنها منبعی برای ارتباطات اجتماعی هستند، بلکه فضایی را فراهم می‌کنند که می‌توان از آن برای شناسایی مشکلات روحی و روان‌شناختی افراد استفاده کرد. در پروژه‌ی حاضر، به بررسی و تشخیص بیماری‌های روحی و روان‌شناختی از جمله افسردگی، استرس، اعتیاد و تمایل به خودکشی از طریق تحلیل محتوای متنی منتشر شده توسط کاربران در شبکه‌های اجتماعی پرداخته شده است. این پروژه بخشی از یک پروژه‌ی بزرگتر با عنوان ”چت‌بات روانشناس“، پروژه ارشد آقای صادق جعفری است که در آن، پیش از ارائه‌ی پاسخ نهایی به کاربر، ورودی‌های او تحلیل و برچسب‌گذاری می‌شود تا مشخص گردد که آیا کاربر با مشکلاتی مانند افسردگی یا استرس روبه‌رو است یا خیر. همانطور که در تصویر ۱-۱ مشاهده می‌کنید، چت‌بات روانشناس از بخش‌های گوناگونی تشکیل شده است و در این پروژه به بخش چهارم که مربوط به تشخیص مشکل و اختلالات متن ورودی است، می‌پردازیم. این پایان‌نامه شامل شش بخش اصلی است. در بخش دوم^۲، به مفاهیم اولیه و پایه‌ای مرتبط با تشخیص بیماری‌های روحی و روش‌های تحلیل متنی پرداخته شده است. در بخش سوم^۳، کارهای مشابه و تحقیقات پیشین در این حوزه بررسی و مقایسه شده‌اند. در بخش چهارم^۴، آزمایشات انجام شده با استفاده از الگوریتم‌های یادگیری ماشین و مدل‌های زبانی بزرگ بر روی مجموعه داده‌های مختلف و آزمایش‌های و نتایج آنها شرح داده

شده است. در این بخش سعی شده است تا مدل‌ها و الگوریتم‌های متنوعی مورد بررسی قرار گیرند، اما به دلیل محدودیت‌های منابع سخت‌افزاری، امکان استفاده از برخی مدل‌ها و داده‌های پیچیده‌تر فراهم نبوده است. بخش پنجم^۵ نیز مربوط به نتیجه‌گیری و ارائه پیشنهادها می‌باشد. در نهایت، مراجع و منابع مورد استفاده در این پژوهش آورده شده‌اند.



شکل ۱-۱: ساختار کلی پروژه چت‌بات

فصل ۲

تعاریف و مفاهیم پایه

۱-۲ مقدمه

در این فصل به تعاریف و مفاهیم پایه موردنیاز در این پایان‌نامه خواهیم پرداخت. ابتدا مفاهیم رایج در هوش مصنوعی و علوم مربوط با آن را بیان می‌کنیم. سپس به تعریف رده‌بندی و بررسی انواع آن اعم از دو برچسبه، چند برچسبه و چند کلاسی می‌پردازیم. پس از آن به بررسی انواع روشهای مورد استفاده، مدل‌های رایج در رده‌بندی متن و الگوریتم‌های رده‌بندی رایج در شناسایی برچسب در داده‌های متنی می‌پردازیم.

۲-۲ تعاریف و مفاهیم پایه

۱-۲-۲ هوش مصنوعی

هوش مصنوعی به توانایی ماشین‌ها و سیستم‌های کامپیوتری برای انجام وظایفی که به طور معمول به هوش انسانی نیاز دارند، اطلاق می‌شود. این وظایف شامل یادگیری، استدلال، حل مسئله، درک زبان طبیعی و حتی تشخیص الگوها می‌باشد.

۲-۲-۲ پردازش زبان طبیعی

پردازش زبان طبیعی شاخه‌ای از هوش مصنوعی است که به تعامل بین کامپیوترها و زبان انسانی می‌پردازد. هدف اصلی این حوزه، توانمندسازی کامپیوترها برای درک، تفسیر و تولید زبان انسانی است.

۲-۲-۳ یادگیری نظارت‌شده

یادگیری نظارت‌شده یک نوع از یادگیری ماشین است که در آن مدل با استفاده از داده‌های برچسب‌دار آموزش داده می‌شود. هدف این نوع یادگیری، پیش‌بینی برچسب‌های خروجی برای داده‌های جدید بر اساس الگوهای یادگرفته‌شده از داده‌های آموزشی است.

۲-۲-۴ استخراج متن

استخراج متن به فرآیند تبدیل متن‌های غیرساختاریافته به فرمت‌های ساختاریافته برای شناسایی الگوها و استخراج اطلاعات معنادار اطلاق می‌شود. این فرآیند شامل تکنیک‌های مختلفی مانند استخراج اطلاعات، تحلیل احساسات و خوشه‌بندی متون است.

۲-۲-۵ پیش‌پردازش متن

پیش‌پردازش متن یکی از مراحل اولیه در پردازش زبان طبیعی است که شامل مجموعه‌ای از تکنیک‌ها برای آماده‌سازی متن خام برای تحلیل‌های بعدی می‌شود. این فرآیند شامل مراحل مختلفی مانند نرمال‌سازی، توکن‌سازی، حذف کلمات توقف، و تبدیل کلمات به ریشه یا لِمّا می‌باشد. هدف اصلی پیش‌پردازش متن، کاهش پیچیدگی و حجم داده‌ها و بهبود دقت مدل‌های یادگیری ماشین است.

لِمّا تیزه کردن

لِمّا تیزه کردن^۱ فرآیندی است که در آن کلمات به شکل پایه یا لِمّا خود تبدیل می‌شوند. به عنوان مثال، کلمات “رفتن”، “رفته”، و “می‌رود” همگی به لِمّا “رفت” تبدیل می‌شوند. این تکنیک به مدل‌های یادگیری

^۱Lemmatization

ماشین کمک می‌کند تا کلمات مختلف با معنای مشابه را به عنوان یک واحد در نظر بگیرند و دقت تحلیل‌ها را افزایش دهند.

ریشه‌یابی

ریشه‌یابی^۲ فرآیندی است که در آن کلمات به ریشه یا استم خود کاهش می‌یابند. برخلاف لمّاتیزه کردن، ریشه‌یابی ممکن است کلمات را به شکلی تبدیل کند که لزوماً یک کلمه معتبر در زبان نباشد. به عنوان مثال، کلمه “رفتن” ممکن است به “رفت” کاهش یابد. این تکنیک به کاهش حجم داده‌ها و افزایش سرعت پردازش کمک می‌کند.

۲-۲-۶ طبقه‌بندی متن

طبقه‌بندی متن فرآیندی است که در آن متون به دسته‌ها یا کلاس‌های مختلف تقسیم می‌شوند. این فرآیند می‌تواند برای تحلیل احساسات، تشخیص موضوعات و حتی ارزیابی صحت گرامری متون استفاده شود.

طبقه‌بندی دودویی

طبقه‌بندی دودویی یکی از انواع الگوریتم‌های یادگیری ماشین است که داده‌ها را به یکی از دو دسته ممکن تقسیم می‌کند. به عنوان مثال، تشخیص ایمیل‌های اسپم و غیر اسپم یا تشخیص بیماری در بیماران.

طبقه‌بندی چندبرچسبی

در طبقه‌بندی چندبرچسبی، هر نمونه می‌تواند به بیش از یک کلاس تعلق داشته باشد. این نوع طبقه‌بندی در مواردی مانند برچسب‌گذاری تصاویر یا دسته‌بندی اسناد با موضوعات متعدد کاربرد دارد.

طبقه‌بندی چندکلاسه

طبقه‌بندی چندکلاسه به فرآیندی اطلاق می‌شود که در آن داده‌ها به بیش از دو کلاس تقسیم می‌شوند. به عنوان مثال، تشخیص نوع حیوانات در تصاویر که می‌تواند شامل سگ، گربه، پرنده و غیره باشد.

²Stemming

۷-۲-۲ ارزیابی مدل‌ها

برای ارزیابی مدل‌ها و مقایسه عملکرد آن‌ها با یکدیگر از روش‌های گوناگونی استفاده می‌شود که در ذیل به ذکر برخی از آن‌ها می‌پردازیم.

صحت

صحت^۳ معیاری است که نشان می‌دهد چه تعداد از پیش‌بینی‌های مدل درست بوده‌اند. فرمول دقت به صورت زیر است:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (۱-۲)$$

دقت

دقت^۴ معیاری است که نشان می‌دهد چه تعداد از پیش‌بینی‌های مثبت مدل واقعاً مثبت بوده‌اند. فرمول دقت به صورت زیر است:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (۲-۲)$$

بازخوانی

بازخوانی^۵ معیاری است که نشان می‌دهد چه تعداد از نمونه‌های مثبت واقعی توسط مدل به درستی شناسایی شده‌اند. فرمول بازخوانی به صورت زیر است:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (۳-۲)$$

^۳Accuracy

^۴Precision

^۵Recall

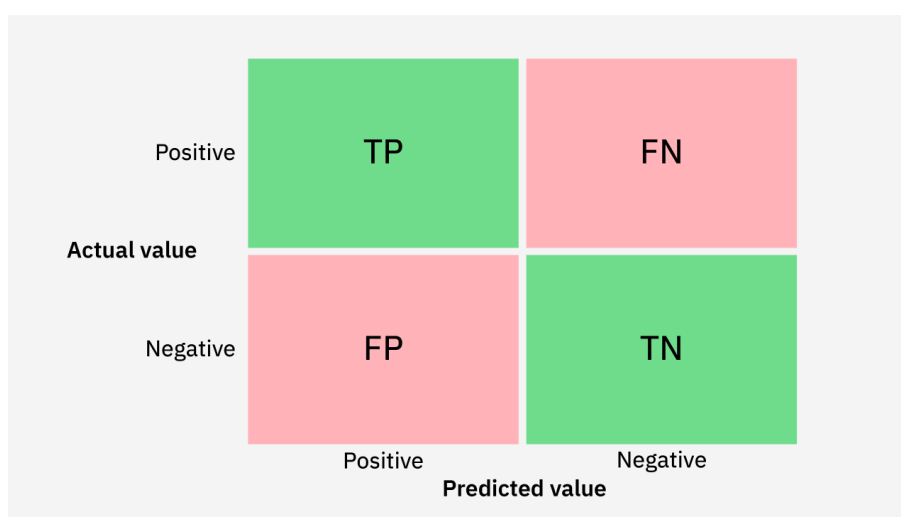
F1

معیار F1 میانگینی هماهنگ از دقت و بازخوانی است و به صورت زیر محاسبه می‌شود:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (۴-۲)$$

ماتریس درهم‌ریختگی

ماتریس درهم‌ریختگی^۶ یک جدول است که عملکرد یک الگوریتم طبقه‌بندی را با مقایسه کلاس‌های واقعی و پیش‌بینی‌شده نشان می‌دهد. این ماتریس شامل چهار مقدار اصلی است: مثبت واقعی (TP)^۷، منفی واقعی (TN)^۸، مثبت کاذب (FP)^۹ و منفی کاذب (FN)^{۱۰} که به تحلیل دقیق‌تر عملکرد مدل کمک می‌کند.



شکل ۲-۱: اضلاع و متغیرهای ماتریس درهم‌ریختگی

^۶Confusion Matrix

^۷True Positive

^۸True Negative

^۹False Positive

^{۱۰}False Negative

امتیاز مساحت زیر منحنی

امتیاز مساحت زیر منحنی (AUC) معیاری است که عملکرد مدل را در تمایز بین کلاس‌های مثبت و منفی اندازه‌گیری می‌کند. این معیار نشان‌دهنده احتمال این است که مدل یک نمونه مثبت را به درستی نسبت به یک نمونه منفی رتبه‌بندی کند.

۲-۲-۸ مدل

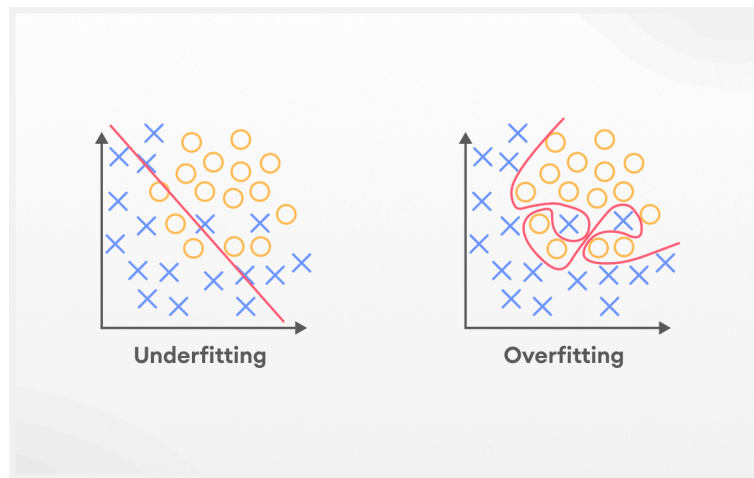
در زمینه یادگیری ماشین و هوش مصنوعی، مدل به یک ساختار ریاضی یا الگوریتم اشاره دارد که برای پیش‌بینی یا تصمیم‌گیری بر اساس داده‌ها استفاده می‌شود. مدل‌ها می‌توانند ساده مانند رگرسیون خطی یا پیچیده مانند شبکه‌های عصبی عمیق باشند. هدف اصلی یک مدل، یادگیری الگوها و روابط موجود در داده‌ها برای انجام وظایف خاص است.

۲-۲-۹ بیش‌برازش

بیش‌برازش (overfitting) زمانی رخ می‌دهد که یک مدل یادگیری ماشین به طور بیش از حد به داده‌های آموزشی خود وابسته می‌شود و الگوهای نویزی و جزئیات غیرمهم را نیز یاد می‌گیرد. این امر باعث می‌شود که مدل در داده‌های جدید عملکرد ضعیفی داشته باشد زیرا به جای یادگیری الگوهای کلی، جزئیات خاص داده‌های آموزشی را حفظ کرده است.

۲-۲-۱۰ کم‌برازش

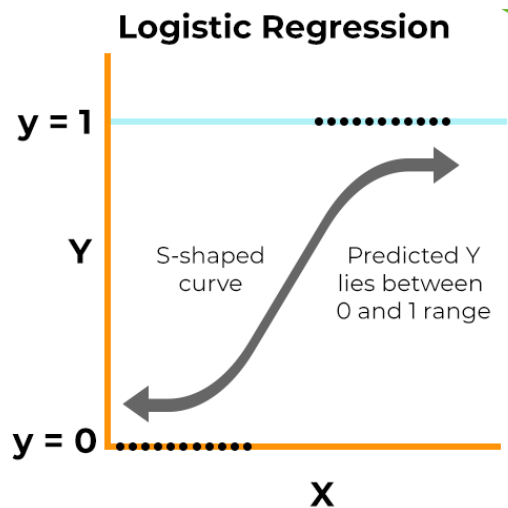
کم‌برازش (underfitting) زمانی رخ می‌دهد که یک مدل یادگیری ماشین نتواند الگوهای موجود در داده‌های آموزشی را به خوبی یاد بگیرد. این مدل‌ها معمولاً بسیار ساده هستند و نمی‌توانند پیچیدگی‌های داده‌ها را به درستی مدل‌سازی کنند، بنابراین در داده‌های آموزشی و داده‌های جدید عملکرد ضعیفی دارند.



شکل ۲-۲: مثالی ساده از وقوع کم‌برازش و بیش‌برازش

۱۱-۲-۲ رگرسیون لجستیک

رگرسیون لجستیک^{۱۱} یک مدل آماری است که برای پیش‌بینی احتمال وقوع یک رویداد باینری (مثلاً بله/خیر یا ۱/۰) استفاده می‌شود. این مدل از تابع لجستیک (sigmoid) برای تبدیل ترکیب خطی متغیرهای مستقل به احتمال استفاده می‌کند.

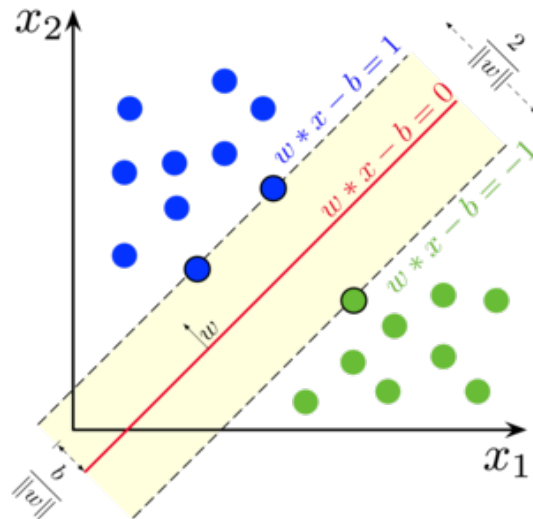


شکل ۲-۳: نمایی ساده از منحنی Logistic Regression

¹¹ Logistic Regression

۱۲-۲-۲ ماشین بردار پشتیبان

ماشین بردار پشتیبان یک الگوریتم یادگیری نظارت شده است که برای طبقه بندی و رگرسیون استفاده می شود. هدف یافتن یک ابرصفحه بهینه در فضای ویژگی ها است که داده ها را به بهترین شکل ممکن جدا کند. این الگوریتم می تواند با استفاده از ترفند کرنل، طبقه بندی غیرخطی را نیز انجام دهد.



شکل ۲-۴: نمایی ساده از نحوه کاربرد و معادله های SVM

۱۳-۲-۲ یادگیری گروهی

یادگیری تجمعی به تکنیکی در یادگیری ماشین اشاره دارد که در آن چندین مدل یادگیری ماشین با هم ترکیب می شوند تا عملکرد کلی بهتری نسبت به هر یک از مدل های منفرد داشته باشند. این روش با ترکیب نتایج چندین مدل، دقت پیش بینی را افزایش می دهد و از خطاهای احتمالی مدل های منفرد جلوگیری می کند.

XGBoost ۱۴-۲-۲

این الگوریتم یک کتابخانه متن باز و بهینه سازی شده برای تقویت گرادیان است که برای حل مسائل مختلف یادگیری ماشین با دقت و سرعت بالا طراحی شده است. این الگوریتم با ترکیب پیش بینی های چندین مدل ضعیف (معمولاً درخت های تصمیم) به صورت تکراری، یک مدل پیش بینی قوی ایجاد می کند.

۲-۲-۱۵ شبکه‌های عصبی مصنوعی

شبکه‌های عصبی مصنوعی الگوریتم‌هایی هستند که از ساختار و عملکرد مغز انسان الهام گرفته‌اند. این شبکه‌ها از لایه‌های مختلفی از نورون‌های مصنوعی تشکیل شده‌اند که به یکدیگر متصل هستند. هر نورون ورودی‌هایی را دریافت می‌کند، آن‌ها را پردازش می‌کند و خروجی‌ای تولید می‌کند که به نورون‌های بعدی منتقل می‌شود. شبکه‌های عصبی برای وظایف مختلفی مانند تشخیص تصویر، پردازش زبان طبیعی و پیش‌بینی سری‌های زمانی استفاده می‌شوند. در بحث شبکه‌های عصبی مفاهیم گوناگونی وجود دارد که در ذیل به ذکر برخی از آن‌ها می‌پردازیم.

زیان

زیان (Loss) یا تابع زیان معیاری است که میزان خطای مدل یادگیری ماشین را اندازه‌گیری می‌کند. این تابع تفاوت بین پیش‌بینی‌های مدل و مقادیر واقعی را محاسبه می‌کند و هدف مدل به حداقل رساندن این مقدار است تا دقت پیش‌بینی‌ها افزایش یابد. روابط برخی از توابع معروف و پر کاربرد در ادامه آورده شده‌اند.

• میانگین مربعات خطا (MSE) Mean Squared Error

$$L_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2-5)$$

• میانگین خطای مطلق (MAE) Mean Absolute Error

$$L_{\text{MAE}} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2-6)$$

• Cross-Entropy

$$L_{\text{CE}} = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (2-7)$$

دوره

دوره (epoch) به یک چرخه کامل از عبور مدل یادگیری ماشین از تمام داده‌های آموزشی اشاره دارد. در هر دوره، مدل یک بار تمام داده‌های آموزشی را می‌بیند و پارامترهای خود را به‌روزرسانی می‌کند. معمولاً مدل‌ها برای چندین دوره آموزش داده می‌شوند تا به دقت مطلوب برسند.

بهینه‌ساز

بهینه‌ساز (optimizer) الگوریتمی است که برای تنظیم پارامترهای مدل یادگیری ماشین به منظور کاهش تابع زیان استفاده می‌شود. بهینه‌سازها با استفاده از گرادیان‌ها، پارامترهای مدل را به گونه‌ای تغییر می‌دهند که تابع زیان به حداقل مقدار خود برسد. از جمله بهینه‌سازها پرکاربرد می‌توان به Adam، AdamW و SGD اشاره کرد.

نرخ یادگیری

نرخ یادگیری (learning rate) یک پارامتر مهم در فرآیند آموزش مدل‌های یادگیری ماشین است که تعیین می‌کند مدل با چه سرعتی پارامترهای خود را به‌روزرسانی کند. نرخ یادگیری بالا می‌تواند باعث نوسانات زیاد در فرآیند آموزش شود، در حالی که نرخ یادگیری پایین ممکن است باعث کندی در همگرایی مدل شود.

کاهش وزن

کاهش وزن (weight decay) یک تکنیک منظم‌سازی است که به منظور جلوگیری از بیش‌برازش در مدل‌های یادگیری ماشین استفاده می‌شود. این تکنیک با اضافه کردن یک جریمه به تابع زیان، از بزرگ شدن بیش از حد وزن‌های مدل جلوگیری می‌کند و به این ترتیب مدل را ساده‌تر و عمومی‌تر می‌سازد.

پارامترهای مدل

پارامترهای مدل شامل وزن‌ها و بایاس‌ها هستند که در طول فرآیند آموزش تنظیم می‌شوند. وزن‌ها نشان‌دهنده اهمیت هر ویژگی ورودی هستند و بایاس‌ها به مدل کمک می‌کنند تا پیش‌بینی‌های دقیق‌تری انجام دهد.

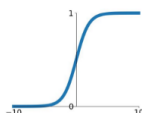
۲-۲-۱۶ تابع فعال‌ساز

تابع فعال‌ساز (Activation Function) در شبکه‌های عصبی مصنوعی، یک تابع ریاضی است که به خروجی یک نورون اعمال می‌شود. هدف اصلی این تابع، معرفی غیرخطی بودن به مدل است تا شبکه بتواند الگوهای پیچیده‌تری را در داده‌ها یاد بگیرد و نمایندگی کند. بدون استفاده از توابع فعال‌ساز، شبکه‌های عصبی تنها قادر به یادگیری روابط خطی ساده خواهند بود. توابع فعال‌ساز مختلفی وجود دارند که هر کدام ویژگی‌ها و کاربردهای خاص خود را دارند، از جمله تابع سیگموئید، تانژانت هایپربولیک، و ReLU.

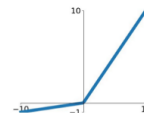
Activation Functions

Sigmoid

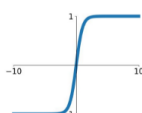
$$\sigma(x) = \frac{1}{1+e^{-x}}$$

**Leaky ReLU**

$$\max(0.1x, x)$$

**tanh**

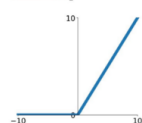
$$\tanh(x)$$

**Maxout**

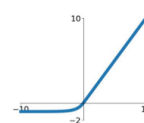
$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ReLU

$$\max(0, x)$$

**ELU**

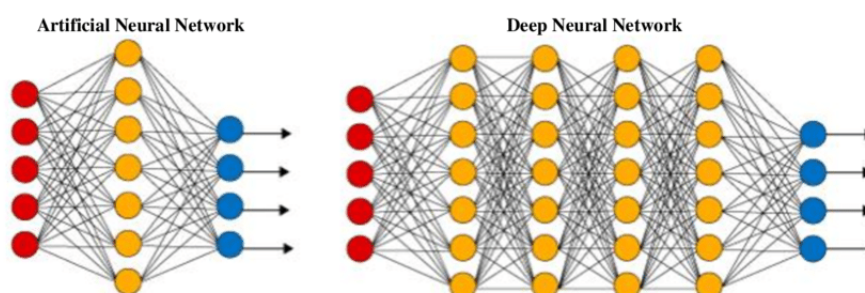
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



شکل ۲-۵: تعدادی از توابع فعال‌ساز در شبکه‌های عصبی به همراه روابط و نمودارها

۲-۲-۱۷ شبکه‌های عصبی عمیق

شبکه‌های عصبی عمیق نوعی از شبکه‌های عصبی هستند که شامل چندین لایه پنهان می‌شوند. این لایه‌های اضافی به شبکه اجازه می‌دهند تا الگوهای پیچیده‌تری را در داده‌ها یاد بگیرد. شبکه‌های عصبی عمیق در بسیاری از کاربردهای پیشرفته مانند تشخیص چهره، ترجمه ماشینی و بازی‌های رایانه‌ای استفاده می‌شوند. این شبکه‌ها به دلیل توانایی‌شان در یادگیری از داده‌های بزرگ و پیچیده، بسیار محبوب شده‌اند.



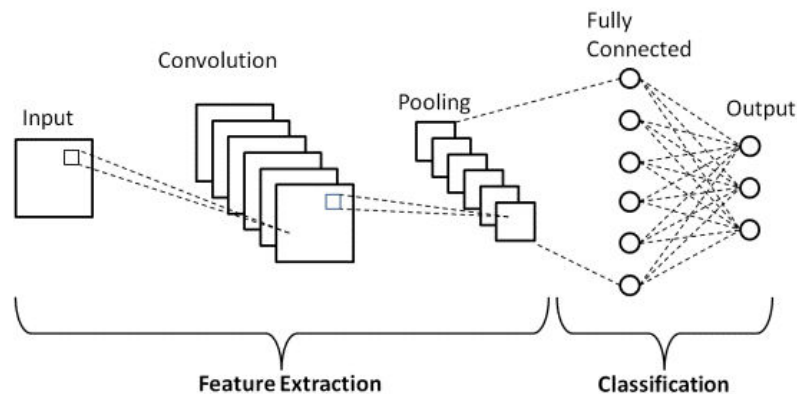
شکل ۲-۶: نمای کلی از شبکه‌های عصبی عمیق

۲-۲-۱۸ شبکه‌های عصبی پیچشی

شبکه‌های عصبی پیچشی^{۱۲}، که به اختصار CNN یا ConvNet نامیده می‌شوند، نوعی از شبکه‌های عصبی عمیق هستند که به طور خاص برای پردازش داده‌های ساختاریافته مانند تصاویر طراحی شده‌اند. این شبکه‌ها از لایه‌های کانولوشن برای استخراج ویژگی‌ها از داده‌ها استفاده می‌کنند. هر لایه کانولوشن شامل فیلترهایی است که به صورت خودکار ویژگی‌های مختلفی مانند لبه‌ها، بافت‌ها و اشکال را از تصاویر استخراج می‌کنند. کاربردها: شبکه‌های عصبی کانولوشنی در بسیاری از کاربردهای عملی مانند تشخیص تصویر، طبقه‌بندی اشیاء، تشخیص چهره، و حتی پردازش زبان طبیعی استفاده می‌شوند. این شبکه‌ها به دلیل توانایی‌شان در یادگیری ویژگی‌های پیچیده از داده‌های بزرگ، در بسیاری از حوزه‌ها بسیار مؤثر هستند.

ساختار: ساختار یک شبکه عصبی کانولوشنی شامل چندین لایه کانولوشن، لایه‌های تجمع (Pooling)، و لایه‌های کاملاً متصل (Fully Connected) است. لایه‌های کانولوشن ویژگی‌های محلی را استخراج می‌کنند، در حالی که لایه‌های تجمع به کاهش ابعاد داده‌ها و افزایش کارایی کمک می‌کنند. لایه‌های کاملاً متصل نیز برای تصمیم‌گیری نهایی استفاده می‌شوند.

¹²Convolutional Neural Networks



شکل ۲-۷: نمای کلی از شبکه‌های CNN

۲-۲-۱۹ شبکه‌های عصبی بازگشتی

شبکه‌های عصبی بازگشتی^{۱۳} نوعی از شبکه‌های عصبی مصنوعی هستند که برای پردازش داده‌های ترتیبی یا سری زمانی طراحی شده‌اند. برخلاف شبکه‌های عصبی پیچشی^{۱۴} که داده‌ها را در یک مرحله پردازش می‌کنند، RNN داده‌ها را در چندین مرحله زمانی پردازش می‌کنند. این شبکه‌ها دارای یک حالت پنهان هستند که در هر مرحله زمانی به‌روزرسانی می‌شود و به شبکه اجازه می‌دهد تا اطلاعات گذشته را به خاطر بسپارد و از آن‌ها در پردازش فعلی استفاده کند.

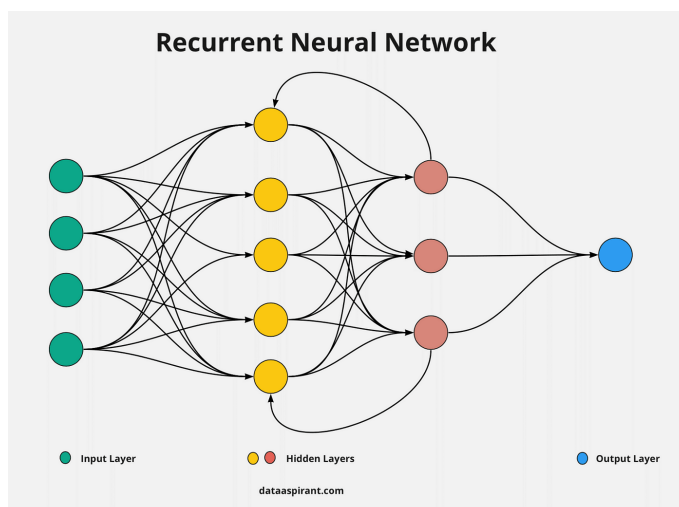
کاربردها: RNN‌ها در بسیاری از کاربردهای عملی مانند پردازش زبان طبیعی، تشخیص گفتار، و پیش‌بینی سری‌های زمانی استفاده می‌شوند. به عنوان مثال، در تحلیل متن، RNN‌ها می‌توانند وابستگی‌های طولانی مدت بین کلمات را یاد بگیرند و درک بهتری از متن ارائه دهند.

۲-۲-۲۰ شبکه‌های حافظه کوتاه‌مدت بلند

LSTM (Long Short Term Memory) نوعی از شبکه‌های عصبی بازگشتی (RNN) است که برای یادگیری وابستگی‌های بلندمدت طراحی شده است. این مدل‌ها به منظور حل مشکلاتی مانند ناپدید شدن و انفجار گرادینان‌ها که در RNN‌های سنتی وجود دارد، معرفی شده‌اند. LSTM‌ها از مکانیزم‌های خاصی مانند حالت سلول و دروازه‌ها برای حفظ اطلاعات در طول توالی‌های بلند استفاده می‌کنند. این مکانیزم‌ها شامل دروازه

¹³ Recurrent Neural Network

¹⁴ Feedforward Neural Networks



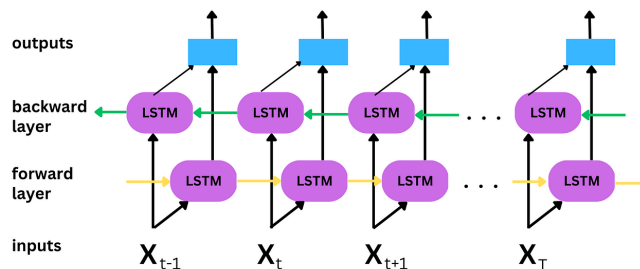
شکل ۲-۸: نمای کلی از شبکه‌های عصبی بازگشتی

فراموشی، دروازه ورودی و دروازه خروجی هستند که به مدل کمک می‌کنند تا تصمیم بگیرد کدام اطلاعات را نگه دارد و کدام را دور بیندازد.

۲-۲-۲۱ شبکه‌های حافظه کوتاه‌مدت بلند دو طرفه

BiLSTM یک توسعه از LSTM سنتی است که شامل دو LSTM موازی است؛ یکی که داده‌ها را از ابتدا به انتها (جهت جلو) پردازش می‌کند و دیگری که داده‌ها را از انتها به ابتدا (جهت عقب) پردازش می‌کند. این ساختار به BiLSTM اجازه می‌دهد تا به هر دو زمینه گذشته و آینده دسترسی داشته باشد، که برای درک کامل‌تر از داده‌های ورودی بسیار مفید است. خروجی‌های هر دو LSTM معمولاً ترکیب می‌شوند تا خروجی نهایی را تشکیل دهند.

مزایا: BiLSTM به دلیل توانایی‌اش در پردازش داده‌ها در هر دو جهت، در وظایفی که نیاز به درک زمینه کامل دارند، مانند تحلیل احساسات و شناسایی موجودیت‌های نام‌برده، عملکرد بهتری دارد. این مدل‌ها معمولاً دقت بیشتری نسبت به LSTM‌های یک‌طرفه دارند.



شکل ۲-۹: نمای کلی از شبکه‌های BiLSTM

۲-۲-۲ مدل‌های زبانی بزرگ

مدل‌های زبانی بزرگ (Large Language Models) یا LLMها، مدل‌های محاسباتی هستند که قادر به تولید زبان یا انجام وظایف پردازش زبان طبیعی هستند. این مدل‌ها با یادگیری روابط آماری از مقادیر عظیمی از متن در طی یک فرآیند آموزش خودنظارتی و نیمه‌نظارتی، این توانایی‌ها را کسب می‌کنند.

ویژگی‌ها و کاربردها: مدل‌های زبانی بزرگ می‌توانند متن را درک و تولید کنند و برای انجام طیف وسیعی از وظایف مانند ترجمه زبان، خلاصه‌سازی متن، پاسخ به سوالات، و حتی تولید محتوای خلاقانه یا کد استفاده می‌شوند. این مدل‌ها به دلیل توانایی‌شان در یادگیری الگوهای پیچیده زبان، در بسیاری از کاربردهای عملی مانند چت‌بات‌ها، دستیارهای مجازی، و تولید محتوا بسیار مؤثر هستند.

معماری و مدل‌های معروف: معماری این مدل‌ها معمولاً بر اساس ترانسفورمرها است که امکان پردازش و تولید داده‌های متنی در مقیاس بزرگ را فراهم می‌کند. برخی از مدل‌های معروف شامل سری GPT از OpenAI، مدل‌های LLaMA از Meta، و مدل‌های BERT و PaLM از Google هستند.

۲-۲-۳ مبدل‌ها

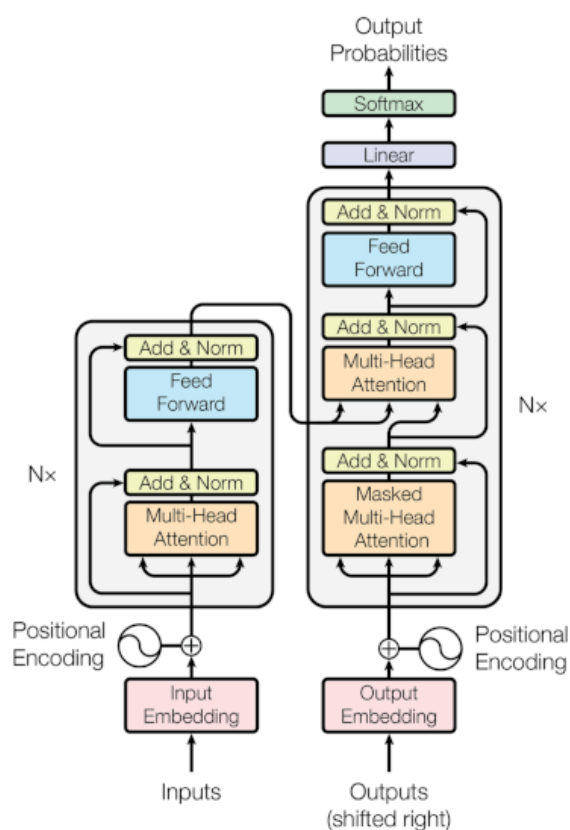
مبدل‌ها^{۱۵} نوعی معماری شبکه عصبی هستند که برای پردازش زبان طبیعی و وظایف مرتبط با آن استفاده می‌شوند. این معماری به دلیل توانایی در مدل‌سازی وابستگی‌های طولانی مدت در متن‌ها بسیار محبوب است. مبدل‌ها را می‌توان به دو دسته رمزگذار و رمزگشا تقسیم نمود. مبدل‌های رمزگذار بطور عمده در مسائل

¹⁵Transformer

رده‌بندی به کار می‌روند. در صورتی که مبدل‌های رمزگشا در تولید متن کاربردهای فراوانی دارند.

۲-۲-۲۴ BERT

خانواده BERT شامل مجموعه‌ای از مدل‌های پردازش زبان طبیعی است که توسط گوگل توسعه یافته‌اند. BERT مخفف “نمایش‌های رمزگذار دوطرفه از مبدل‌ها” است و از معماری رمزگذار مبدل برای پردازش هر توکن از متن ورودی در زمینه کامل تمام توکن‌های قبل و بعد استفاده می‌کند. این مدل‌ها معمولاً بر روی یک مجموعه بزرگ از متون پیش‌آموزش داده می‌شوند و سپس برای وظایف خاص بهینه‌سازی می‌شوند. خانواده BERT شامل مدل‌های مختلفی مانند RoBERTa، ALBERT و DistilBERT است. هر یک از این مدل‌ها بهبودهایی در زمینه‌های مختلف مانند سرعت پردازش، دقت، و کارایی دارند. به عنوان مثال، RoBERTa با استفاده از تکنیک‌های پیشرفته‌تر پیش‌آموزش، دقت بیشتری را ارائه می‌دهد، در حالی که ALBERT با کاهش تعداد پارامترها، کارایی بیشتری دارد.



شکل ۲-۱۰: معماری رمزگذار BERT

۲-۲-۲۵. تنظیم دقیق

تنظیم دقیق^{۱۶} فرآیندی است که در آن یک مدل پیش‌آموزش داده شده برای یک وظیفه خاص بهینه‌سازی می‌شود. در این فرآیند، مدل با استفاده از داده‌های خاص آن وظیفه مجدداً آموزش داده می‌شود تا عملکرد بهتری داشته باشد. به عنوان مثال، یک مدل BERT که برای پردازش زبان طبیعی پیش‌آموزش داده شده است، می‌تواند با استفاده از داده‌های خاص یک وظیفه مانند طبقه‌بندی متون یا ترجمه، تنظیم دقیق شود.

¹⁶Fine tuning

فصل ۳

مروری بر کارهای مرتبط

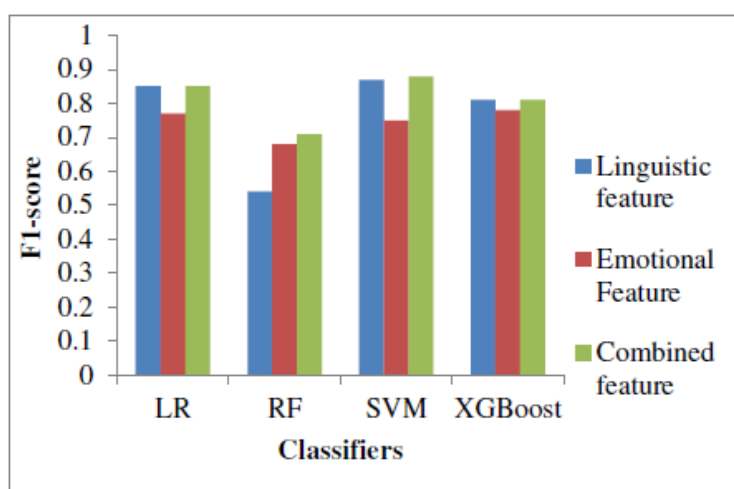
۳-۱ مقدمه

در این فصل به بررسی برخی پژوهش‌های مرتبط انجام شده در گذشته، بررسی مجموعه دادگان آن‌ها و روش‌های مورد استفاده در آن‌ها می‌پردازیم. همچنین به دو چالش **eRisk** و **LT-EDI** که در کنفرانس‌های بین‌المللی Empirical Methods in Natural Language و Conference and Labs of the Evaluation Forum (CLEF) Processing (EMNLP)، به ترتیب برگزار می‌شوند به همراه برخی کارهایی که به تازگی در آن‌ها منتشر شده‌اند، اشاره می‌کنیم. با توجه به پژوهش‌های جامع انجام گرفته، مدل‌ها و الگوریتم‌های رده‌بندی متن را به دو دسته ”براساس ویژگی” و ”عمیق” می‌توانیم تقسیم کنیم [۸، ۱۵، ۱۴].

۳-۲ کارهای پیشین

۳-۲-۱ تشخیص و تحلیل با استفاده از الگوریتم‌های استخراج ویژگی در یادگیری ماشین
کامر و همکاران در پژوهش [۱۲] به تشخیص افسردگی از طریق تحلیل پست‌های کاربران در توییتر با استفاده از تکنیک‌های یادگیری ماشین و پردازش زبان طبیعی می‌پردازد. داده‌ها شامل توییت‌های کاربران افسرده و غیر افسرده است که با استفاده از API‌های توییتر جمع‌آوری شده‌اند. ویژگی‌های استخراج شده شامل تحلیل

احساسات، مدل‌سازی موضوعی با استفاده از LDA^۱، و مدل‌سازی زبانی با استفاده از N-gram و TF-IDF هستند. برای تشخیص افسردگی، چندین مدل یادگیری ماشین از جمله رگرسیون لجستیک^۲، جنگل تصادفی، SVM و XGBoost استفاده شده است. نتایج نشان داد که مدل SVM با ترکیب ویژگی‌های TF- + Trigram + IDF^۳ تحلیل احساسات و تحلیل ایموجی‌ها، به دقت ۸۹٪ دست یافت. این تحقیق اهمیت انتخاب صحیح ویژگی‌ها و ترکیب آن‌ها برای بهبود دقت تشخیص افسردگی در شبکه‌های اجتماعی را نشان می‌دهد. تصویر ۳-۵ خروجی رده‌بندی‌های گوناگون را با توجه به معیار F1 نمایش می‌دهد.



شکل ۳-۱: مقدار امتیاز F1 محاسبه شده بر روی مجموعه داده توسط رده‌بندی‌های گوناگون در پژوهش [۱۲]

¹Linear discriminant analysis

²Logistic Regression

³Term Frequency - Inverse Document Frequency

در مقاله ویلاپرز و همکاران [۲۳]، محققان از داده‌های توییت برای تشخیص خودکار نه نوع اختلال روانی استفاده کرده‌اند. آنها دو مجموعه داده [۲۴] به زبان‌های انگلیسی و اسپانیایی ایجاد کردند. مجموعه "تشخیص داده شده" شامل توییت‌های ۱۵۰۰ کاربر است که در پست‌های خود به تشخیص یکی از اختلالات روانی اشاره کرده‌اند. این مجموعه داده از نوع چندکلاسه می‌باشد و نام کلاس‌های آن به شرح زیر می‌باشند:

● Anxiety

● Bipolar Disorder

● Depression Severity

● Eating Disorder

● Obsessive-Compulsive Disorder (OCD)

● Post-Traumatic Stress Disorder (PTSD)

● Schizophrenia

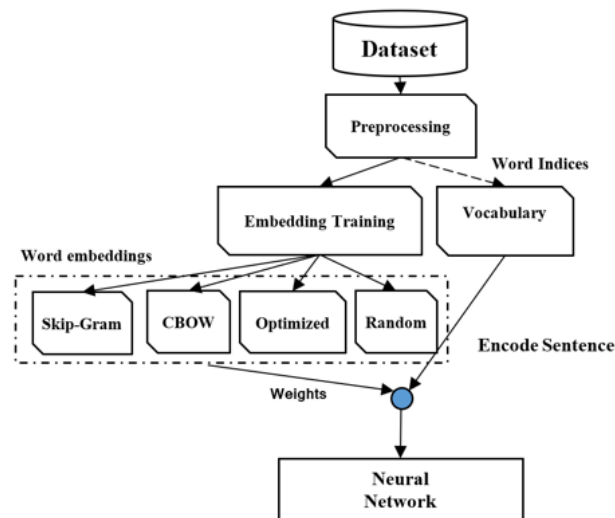
● CONTROL

● Autism Spectrum Disorder (ASD)

● Attention Deficit Hyperactivity Disorder (ADHD)

برای تحلیل داده‌ها، محققان از ویژگی‌های متنی مختلفی مانند n-grams، q-grams، برچسب‌های بخش‌های گفتار (POS)، مدل‌سازی موضوعی، LIWC و تعبیه‌های کلمات استفاده کردند. آنها مدل‌های مختلف یادگیری ماشین و یادگیری عمیق را برای دو وظیفه طبقه‌بندی باینری (تشخیص داده شده در مقابل غیر تشخیص داده شده) و طبقه‌بندی چندکلاسه (تشخیص نوع خاص اختلال) آموزش دادند. مدل‌های XGBoost و شبکه عصبی کانولوشن (CNN) بهترین عملکرد را در هر دو وظیفه داشتند. در طبقه‌بندی باینری، مدل‌های آنها به ترتیب AUC ۰.۸۳۵ و ۰.۸۴۶ را برای مجموعه داده‌های توییت اسپانیایی و انگلیسی به دست آوردند. در طبقه‌بندی چندکلاسه، AUC به ترتیب ۰.۷۱۲ و ۰.۶۹۷ بود.

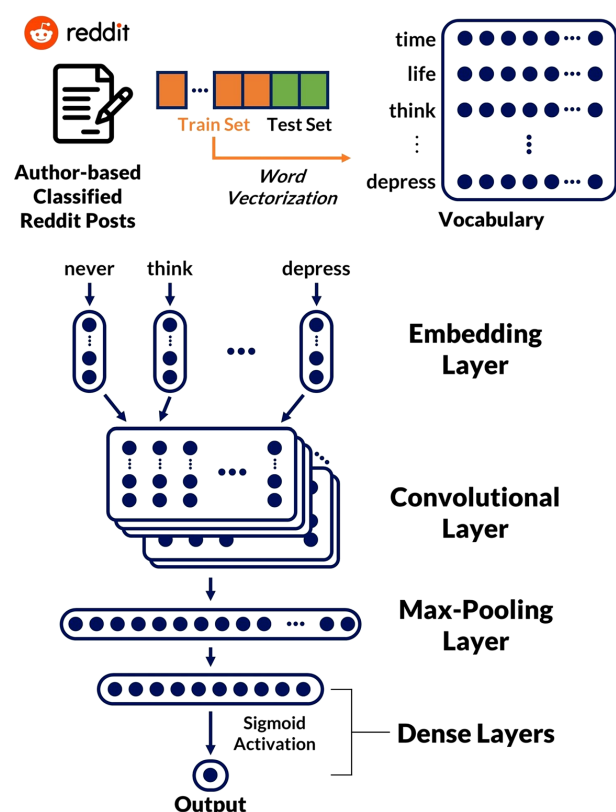
در اورابی و همکاران [۲۳] به شناسایی افسردگی در کاربران توییتر با استفاده از مدل‌های عمیق یادگیری ماشین پرداخته شده است. داده‌های مورد استفاده شامل مجموعه CLPsych 2015 و Bell Let's Talk است که شامل پست‌های کاربران افسرده، افراد عادی و مبتلایان به PTSD است. مدل‌های مورد استفاده شامل سه مدل مبتنی بر شبکه‌های عصبی پیچشی (CNN) و یک مدل مبتنی بر شبکه‌های عصبی بازگشتی (RNN) با لایه‌های LSTM دوطرفه است. بهترین نتیجه با مدل CNNWithMax و بهینه‌سازی تعبیه‌سازی کلمات به دست آمد که به دقت ۸۷.۹۵٪ و مساحت زیر منحنی^۴ برابر با ۰.۹۵۱ در مجموعه داده CLPsych2015 رسید. این مدل‌ها توانستند با عملکرد خوب خود، قابلیت تعمیم‌پذیری بر روی داده‌های دیده‌نشده را نشان دهند.



شکل ۲-۳: معماری سیستم در پژوهش [۲۳]

و در مقاله دیگری توسط کیم و همکاران [۱۱]، به تحلیل محتوای کاربران شبکه Reddit مبتلا به بیماری‌های روانی همچون افسردگی و اضطراب از طریق مدل‌های یادگیری عمیق پرداخته شده است. تصویر ۳-۳ معماری شبکه CNN رده بند متن را نمایش می‌دهد.

⁴Area Under Curve



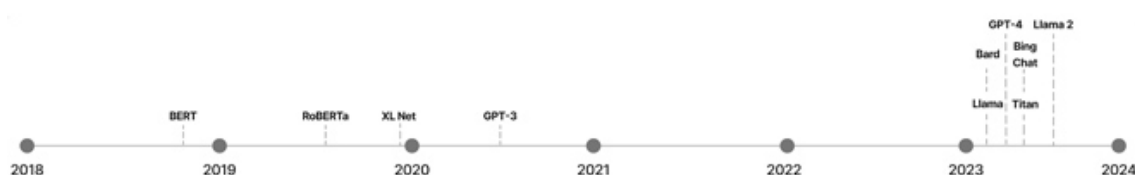
شکل ۳-۳: معماری شبکه CNN در پژوهش [۱۱]

۳-۲-۲ تشخیص و تحلیل با استفاده از مدل‌های زبانی بزرگ

مدل‌های زبانی بزرگ به عنوان ابزارهای پیشرفته پردازش زبان طبیعی توانسته‌اند تحول چشمگیری در زمینه‌های مختلف از جمله رده‌بندی متن ایجاد کنند. این مدل‌ها، که بر پایه معماری‌های پیچیده‌ای نظیر مدل‌ها توسعه یافته‌اند، قابلیت تحلیل و درک متن را به صورت عمیق‌تری نسبت به روش‌های سنتی فراهم می‌آورند [۱۶، ۲۵]. یکی از کاربردهای مهم این مدل‌ها در رده‌بندی متن است که شامل تخصیص دسته‌بندی‌های مناسب به متن‌های ورودی بر اساس محتوای آن‌ها می‌شود. تحقیقات اخیر نشان داده‌اند که مدل‌های زبانی بزرگ مانند GPT-4 و BERT [۵] قادر به دستیابی به دقت‌های بالا و عملکرد بهینه در این زمینه هستند. همچنین مقاله ارزیابی [۷] گزارشی جامع از تاریخچه ساخت مدل‌های رمزگذار و رمزگشا، کاربردها و چالش‌های آن‌ها در طبقه‌بندی متون ارائه می‌کند.

این پیشرفت‌ها، توانایی مدل‌های زبانی بزرگ را در مقابله با چالش‌های پیچیده مرتبط با رده‌بندی متن، از جمله درک مفاهیم ضمنی و معانی نهاده شده، بهبود بخشیده است. نتایج تجربی و مقایسه‌های صورت گرفته

در مطالعات مختلف نشان داده‌اند که این مدل‌ها به طور چشمگیری نسبت به روش‌های سنتی مانند مدل‌های مبتنی بر ویژگی‌های دستی، عملکرد بهتری دارند. این روند نشان‌دهنده اهمیت و تاثیر فزاینده مدل‌های زبانی بزرگ در دنیای پردازش زبان طبیعی و کاربردهای آن در رده‌بندی متن است.



شکل ۳-۴: سیر زمانی و تاریخچه انتشار مبدل‌ها [۷]

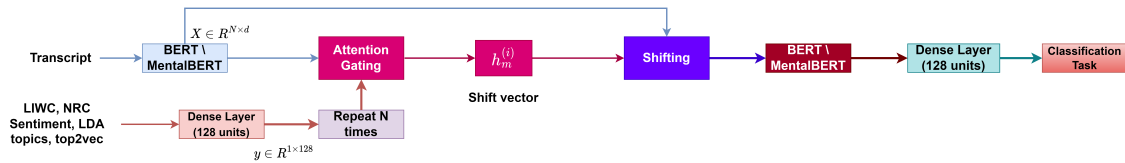
ایلیاس و همکاران در مقاله [۱۰] به بررسی روش‌های مبتنی بر مدل‌های مبدل برای تشخیص استرس و افسردگی از داده‌های شبکه‌های اجتماعی پرداخته شده است. مدل‌های مورد استفاده شامل BERT و MentalBERT هستند که با استفاده از اطلاعات زبانی اضافی بهبود یافته‌اند. این اطلاعات شامل ویژگی‌های احساسی NRC، LIWC، مدل‌سازی موضوعی با LDA، و Top2Vec است.

● Dreaddit برای تشخیص متن استرس‌زا و غیر استرس‌زا

● Depression Mixed برای تشخیص پست‌های افسرده و غیر افسرده

● Depression Severity برای تخمین شدت افسردگی (کم، متوسط، شدید)

ادغام این ویژگی‌های زبانی با مدل‌های ترانسفورمر دقت تشخیص را بهبود می‌بخشد. بهترین مدل با استفاده از MentalBERT و ویژگی‌های Top2Vec به دقت ۹۳.۰۶٪ در تشخیص افسردگی و ۸۱.۱۲٪ در تشخیص استرس رسید. همچنین، استفاده از روش Label Smoothing به بهبود کالبراسیون مدل‌ها کمک کرده است.



شکل ۳-۵: معماری سیستم در پژوهش [۱۰]

۳-۲-۳ چالش eRisk

چالش eRisk که بخشی از کنفرانس و آزمایشگاه‌های ارزیابی CLEF است، بر شناسایی زود هنگام خطرات از طریق نوشته‌های آنلاین کاربران، با تمرکز بر مسائل مربوط به سلامت و ایمنی، تمرکز دارد. هدف اصلی این چالش توسعه مدل‌هایی است که قادر به تشخیص علائم اولیه مشکلات سلامت روان از محتوای شبکه‌های اجتماعی باشند. در نسخه ۲۰۲۳ این چالش، سه وظیفه اصلی وجود داشت:

- تشخیص علائم افسردگی: رتبه‌بندی جملات از نوشته‌های کاربران بر اساس ارتباط آن‌ها با ۲۱ علامت افسردگی ذکر شده در پرسشنامه BDI-II.
- تشخیص قماربازی بیمارگونه: شناسایی علائم اولیه قماربازی بیمارگونه با تحلیل پی‌درپی پست‌های کاربران و تصمیم‌گیری به موقع درباره اینکه آیا کاربر نشانه‌هایی از اعتیاد به قمار نشان می‌دهد یا خیر.
- تخمین شدت اختلالات خوردن: به‌طور خودکار پاسخ‌های کاربران به پرسشنامه مربوط به اختلالات خوردن را از طریق تحلیل نوشته‌های آن‌ها در شبکه‌های اجتماعی تخمین زده و شدت علائم مختلف مرتبط با اختلالات خوردن را ارزیابی می‌کند.

تیم UNSL [۳۲]، در هر دو وظیفه ۱ و ۲ شرکت کرد. برای وظیفه ۱، آن‌ها سه پیشنهاد ارائه دادند. دو پیشنهاد اول بر شباهت بردارهای تعبیه‌شده متنی بین پست‌ها و علائم پرسشنامه افسردگی بک تمرکز داشتند. پیشنهاد سوم از یک استراتژی پرسش‌گذاری استفاده کرد و وظیفه بازبازی جملات را به یک مسئله زبان ماسک‌شده تبدیل کرد. آن‌ها از ChatGPT برای تولید مثال‌های مصنوعی مرتبط با یک علامت خاص استفاده کردند و مدل RoBERTa را برای حل مسئله زبان ماسک‌شده آموزش دادند. استراتژی مبتنی بر پرسش‌گذاری نتایج بهتری نسبت به پیشنهادات مبتنی بر شباهت ارائه داد و نتایج امیدوارکننده‌ای برای بازبازی اطلاعات به دست آورد. برای وظیفه ۲، که مربوط به تشخیص زود هنگام نشانه‌های قمار پاتولوژیک بود، تیم UNSL سه مدل

بهینه‌سازی شده پیشنهاد داد که به دنبال آن یک سیاست تصمیم‌گیری بر اساس معیارهای تعریف شده توسط یک چارچوب تشخیص زودهنگام اعمال شد. یکی از مدل‌ها شامل یک واژگان گسترده با کلمات مهم خاص دامنه بود.

تیم Xabi_EHU [۱۳]، در وظیفه ۲ شرکت کرد. برای حل مشکل عدم تعادل کلاس‌ها، جایی که گروه قماربازان پاتولوژیک کوچکتر از گروه کنترل بود، تیم از یک شبکه عصبی با تابع زیان سفارشی استفاده کرد. این رویکرد به آنها اجازه داد تا جریمه‌های مثبت و منفی کاذب را تنظیم کنند و انعطاف‌پذیری برای برآورده کردن نیازهای خاص فراهم کنند. با استفاده از این تابع زیان سفارشی، رویکرد آموزشی آنها جریمه‌های مثبت و منفی کاذب را به طور مساوی اعمال نکرد.

تیم RiskBuster [۹]، در وظیفه ۳ شرکت کرد. رویکرد آنها شامل استفاده از یک روش مدل‌سازی موضوع مبتنی بر ترانسفورمر برای اندازه‌گیری شدت علائم اختلالات خوردن بود. تیم، چارچوب BERTopic را سفارشی کرد تا توزیع موضوعات را در سطح کاربر به دست آورد. این توزیع‌های موضوعی سپس به عنوان ویژگی‌های ورودی برای وظایف طبقه‌بندی پایین‌دستی استفاده شدند. با استفاده از مدل‌سازی موضوع، تیم هدف داشت تا تم‌های زیرین و شدت علائم اختلالات خوردن را در نوشته‌های رسانه‌های اجتماعی شناسایی کند. برای بهبود کیفیت تعبیه‌ها، تیم RiskBuster مدل زبان مبتنی بر مبدل MentalBERT را به دامنه اختلالات خوردن تطبیق داد.

تیم uOttawa [۳۵]، در وظیفه اول شرکت کرد. آن‌ها از ترکیبی از مدل‌های تعبیه کلمات شامل GloVe، RoBERTa، DistilBERT و Universal Sentence Encoder (USE) استفاده کردند. از میان مدل‌های ارائه‌شده توسط این تیم، USE به‌طور مداوم در تمام معیارها بهترین عملکرد را داشت، چه با استفاده از رأی‌گیری اکثریت و چه با استفاده از اتفاق نظر برای ارزیابی نتایج. روش‌های مبتنی بر نمایش‌های متنی، از جمله USE، RoBERTa و DistilBERT، عملکرد بهتری نسبت به روش‌های مبتنی بر تعبیه‌های GloVe داشتند. پس از تحلیل تجربی، تیم فرض کرد که عملکرد بهتر روش‌های مبتنی بر نمایش‌های متنی ممکن است به دلیل حذف ضمایری باشد که می‌توانند اطلاعات مهمی در گفتمان ارائه دهند. این حذف بر روی روش‌های مبتنی بر GloVe اعمال نشده بود.

تیم UMU [۲۴]، استراتژی‌های مختلفی را برای سه وظیفه متمایز در این چالش معرفی کرد. برای وظیفه اول تیم آن را به عنوان یک مسئله پرسش و پاسخ در نظر گرفت و دو پیشنهاد ارائه داد: یکی برای متن‌های تک‌زبانه و دیگری برای متن‌های چندزبانه. آن‌ها از مدل "مبدل جملات از پیش آموزش‌دیده" برای ارزیابی

شدت هر علامت افسردگی در مجموعه متون کاربران استفاده کردند. تیم از یک روش مبتنی بر لغت‌نامه افسردگی برای انتخاب متون مرتبط با افسردگی و رتبه‌بندی ۲۱ علامت افسردگی پرسشنامه BDI استفاده کرد. با در نظر گرفتن وظیفه اول به عنوان مسئله پرسش و پاسخ، آن‌ها سوالات پرسشنامه BDI را به پاسخ‌های ممکن در نوشته‌های کاربران متصل کردند تا درجه ارتباط هر متن با علائم افسردگی را تعیین کنند. برای وظیفه دوم، که بر شناسایی زودهنگام علائم قماربازی بیمارگونه متمرکز بود، تیم UMU پنج اجرای مختلف ارائه داد که ترکیبی از استراتژی‌های تصمیم‌گیری با مدل‌های طبقه‌بندی بود. در وظیفه سوم، تیم با استفاده از "مبدل جملات از پیش آموزش دیده" و استخراج ویژگی‌های ایموجی برای بهبود عملکرد مدل، یک اجرا ارائه داد.

تیم SINAI [۱۸]، در وظیفه دوم شرکت کرد. آن‌ها روش‌های مختلفی را با استفاده از مدل‌های ترانسفورمر XLM-RoBERTa و RoBERTa توسعه دادند و نوآورانه‌ترین پیشنهاد آن‌ها ترکیبی از معماری‌های LSTM و RoBERTa بود. روش آن‌ها شامل کدگذاری پست‌های کاربران با استفاده از RoBERTa برای به دست آوردن تعبیه‌ها (embeddings) بود که سپس از طریق LSTM عبور داده شدند. پیش‌بینی نهایی توسط یک شبکه پیش‌خور (Feed Forward Network) انجام شد. آن‌ها همچنین مراحل پیش‌پردازش را برای آماده‌سازی داده‌ها قبل از آموزش طراحی کردند. با این رویکرد، تیم بالاترین امتیاز Recall را برای وظیفه دودویی به دست آورد و دومین مقدار ERDE را ثبت کرد. از نظر ارزیابی رتبه‌بندی نیز آن‌ها یکی از بالاترین جایگاه‌ها را کسب کردند. تیم RELAI [۲۰]، در وظیفه اول و وظیفه دوم شرکت کرد. برای وظیفه اول، تیم از انکدرهای جملات مبتنی بر مبدل استفاده کرد و از روش Okapi BM25 برای انتخاب جملات بهره برد. جملات انتخاب شده سپس وارد یک وظیفه تشابه‌سنجی شدند و امتیازات تشابه برای رتبه‌بندی با استراتژی‌های مختلف مورد استفاده قرار گرفتند. اگرچه نتایج بهبود قابل توجهی نداشتند، اما در برخی موارد از ساده‌ترین خط‌مبنای خود بهتر عمل کردند. در وظیفه دوم، تیم RELAI از رویکردهای "سبک" استفاده کرد. یک روش شامل استخراج ویژگی‌های استیلومتری و سطحی مانند فراوانی کاراکترها و n-gram و طول جملات بود. این ویژگی‌ها سپس به یک پرسبترون چندلایه برای طبقه‌بندی داده شدند. در روش دیگر، از مدل‌سازی موضوعی پست‌ها استفاده شد و موضوعات استخراج شده به عنوان ورودی به پرسبترون چندلایه داده شدند. در رویکرد نهایی، از یک روش احتمالاتی استفاده شد که با تخمین نسبت کاربران مثبت از نسخه‌های قبلی، توزیع گامای مربوط به آن‌ها را تعیین می‌کرد.

تیم BFH-AMI [۲۱]، وابسته به هوش ماشین کاربردی سوئیس، در وظیفه سوم چالش شرکت کرد. این

تیم از یک مدل رگرسیون لجستیک استفاده کرد که شامل نمایه‌های کاربر و پرسش بود که از مدل‌های زبانی بزرگ استخراج شده بودند. برای پرداختن به چالش، نویسندگان از دو مدل مختلف، یعنی BERT و GPT-Large، برای تولید نمایه‌هایی برای پست‌ها و انتشارات کاربران استفاده کردند. این نمایه‌ها سپس به عنوان ویژگی‌های ورودی برای مدل رگرسیون لجستیک به کار رفتند. هدف این بود که الگوها و اطلاعات زیرین در تاریخچه نوشتاری رسانه‌های اجتماعی بیماران را که می‌تواند به پیش‌بینی پاسخ‌های آن‌ها به EDE-Q کمک کند، شناسایی کنند. استفاده از رگرسیون لجستیک به تیم این امکان را داد که یک مدل پیش‌بینی‌کننده بسازد که می‌توانست احتمال پاسخ‌های مختلف را بر اساس ویژگی‌های استخراج شده برآورد کند. با گنجاندن نمایه‌های کاربر و پرسش، مدل می‌توانست به طور مؤثری از اطلاعات معنایی و زمینه‌ای در نوشتارهای رسانه‌های اجتماعی استفاده کند.

تیم BLUE [۴]، در وظیفه اول شرکت کرد. رویکرد آن‌ها شامل استفاده از بازیاب‌های متراکم برای جستجوی معنایی بود، با هدف بازیابی جملات مرتبط از یک مجموعه. آن‌ها از دو نوع پرسش برای فرآیند جستجو استفاده کردند. ابتدا، توصیف‌های هر علامت موجود در پرسشنامه BDI-II را به عنوان پرسش‌ها به کار بردند. این توصیف‌های علامت به عنوان پرسش‌ها برای جستجو عمل کردند. علاوه بر این، آن‌ها از ChatGPT، یک مدل زبانی، برای تولید پرسش‌های مصنوعی برای هر علامت در پرسشنامه BDI-II بهره بردند. تیم معتقد بود که استفاده از ChatGPT برای تولید پرسش‌های مصنوعی تنوع بیشتری در عبارات ایجاد خواهد کرد و احتمالاً جملات مرتبط‌تری در طول فرآیند جستجو به دست خواهد آمد. تیم پنج اجرای مختلف از رویکرد خود را انجام داد و از دو مدل مبتنی بر ترنسفورمر برای نمایه‌سازی پست‌های رسانه‌های اجتماعی استفاده کرد. این مدل‌ها شامل پاسخ‌های اصلی و تولید شده BDI-II، MentalRoBERTa و یک نوع MPNet بودند. به‌طور شگفت‌انگیزی، مدلی که از پاسخ‌های اولیه پرسشنامه BDI-II به عنوان پرسش‌ها برای جستجوی معنایی استفاده کرد، عملکرد بهتری نسبت به مدلی که از پرسش‌های تولید شده استفاده کرد، داشت. همچنین مشاهده شد که داده‌های مصنوعی تولید شده برای این وظیفه خاص بسیار خاص بودند.

تیم BioNLP-IISERB [۳۱]، در وظیفه دوم چالش شرکت کرد. این تیم پنج اجرای مختلف ارسال کرد و از چارچوب‌ها و استراتژی‌های مختلف داده‌کاوی برای طبقه‌بندی متن و مهندسی ویژگی‌ها استفاده کرد. برای استخراج ویژگی‌ها از پست‌های رسانه‌های اجتماعی، BioNLP-IISERB از رویکردهای سنتی و مبتنی بر ترنسفورمر بهره برد. آن‌ها از مدل کیسه کلمات، به‌ویژه وزن‌دهی TF-IDF، و همچنین نمایه‌های مبتنی بر ترنسفورمر مانند BERT و Longformer استفاده کردند. در مرحله طبقه‌بندی، تیم چندین طبقه‌بند

را مورد بررسی قرار داد تا اثرگذاری آن‌ها در این وظیفه را تعیین کند. آن‌ها از طبقه‌بندهایی مانند تقویت تطبیقی (AdaBoost)، رگرسیون لجستیک، ماشین بردار پشتیبان، درخت تصمیم، جنگل تصادفی و طبقه‌بندهای مبتنی بر ترنسفورمر استفاده کردند. در میان رویکردهای مختلف خود، BioNLP-IISERB متوجه شد که ترکیب ماشین بردار پشتیبان و طبقه‌بند تقویت تطبیقی بهترین عملکرد را برای آن‌ها به ارمغان آورده است.

تیم ELiRF-UPV [۲۲]، در وظیفه دوم چالش شرکت کرد. رویکرد آن‌ها بر استفاده از SVMها، یک تکنیک یادگیری ماشین سنتی، متمرکز بود. تیم ELiRF-UPV در آزمایشات خود از SVMها با هسته‌ها و پارامترهای منظم‌سازی مختلف استفاده کرد. آن‌ها از تکنیک‌های اعتبارسنجی متقابل برای ارزیابی عملکرد طبقه‌بندهای SVM استفاده کردند. برای نمایش داده‌های متنی، از یک وکتورایزر TF-IDF با حداکثر ۵۰۰۰ ویژگی استفاده کردند. به‌ویژه، تیم ELiRF-UPV بالاترین نمرات را در دقت، نمره F1، ERDE5، ERDE50 و F1 وزندهی شده به تأخیر در وظیفه دوم چالش کسب کرد.

تیم Formula-ML [۲۷]، در وظیفه اول شرکت کرد. برای نمایش جملات، Formula-ML از تکنیک‌های مختلفی مانند word2vec، مدل TF-IDF و ترنسفورمرهای جمله استفاده کرد. این تکنیک‌ها به آن‌ها این امکان را داد که داده‌های متنی را به نمایش‌های عددی یا نمایه‌ها تبدیل کنند. برای تعیین درجه همبستگی بین جملات و پاسخ‌ها از پرسشنامه BDI، Formula-ML از یک معیار شباهت کسینوسی نرم استفاده کرد. این معیار ارتباط موضوعات مطرح‌شده در جملات با علائم افسردگی را ارزیابی کرد. نتایج به‌دست‌آمده توسط Formula-ML بسته به مدل نمایه‌سازی مورد استفاده متفاوت بود. مدل‌های مبتنی بر SentenceTransformers عملکرد بسیار خوبی داشتند و حتی در برخی از معیارهای ارزیابی رتبه بالایی کسب کردند. با این حال، مدل‌های مبتنی بر word2vec به همین سطح از عملکرد نرسیدند.

رویکرد تیم Mason-NLP [۲۸]، شامل دو مدل یادگیری عمیق، MentalBERT و RoBERTa، به همراه LSTM بود. تیم چندین مرحله را برای کاهش تعداد جملات قبل از عبور از مدل MentalBERT انجام داد، که نمایه‌ای برای هر جمله محاسبه کرد. سپس، رتبه‌بندی هر علامت با محاسبه شباهت کسینوسی بین نمایه هر جمله و نمایه گزینه‌های علامت تعیین شد. با این حال، نتایج ارزیابی رویکرد آن‌ها با انتظارات نویسندگان مطابقت نداشت. با وجود این نتیجه، آن‌ها فرصت بهبود را پذیرفتند و چندین مسیر برای کارهای آینده پیشنهاد کردند. یکی از پیشنهادها، تنظیم دقیق MentalBERT با استفاده از داده‌های آموزشی BDI بود که می‌توانست به بهبود عملکرد مدل کمک کند.

تیم NailP [۳]، در وظیفه اول شرکت کرد. رویکرد آن‌ها بر روی پیش‌پردازش داده‌ها و محاسبه شباهت

بین نمایش‌های متنی متمرکز بود. در ابتدا، نویسندگان با انتخاب انتشاراتی که محتوای خودارجاعی داشتند، به پیش‌پردازش داده‌ها پرداختند و به‌طور خاص به ضمائر شخصی اشاره کردند. سپس جملات مثبت یا خنثی را فیلتر کردند و فقط بر روی جملاتی با احساس منفی بالقوه تمرکز کردند. نمایه‌های جملات با استفاده از SBERT (Sentence-BERT)، مدلی که نمایه‌های جملات زمینه‌دار تولید می‌کند، محاسبه شد. علاوه بر این، تیم نمایه‌هایی از توصیف‌های اقلام BDI-II با استفاده از همان مدل SBERT به‌دست آورد. برای رتبه‌بندی جملات، یک جستجوی معنایی با فیلترهای مختلف برای شامل یا حذف پست‌های منفی انجام شد.

تیم NLP-UNED-2 [۶]، در وظیفه دوم چالش شرکت کرد. رویکرد آن‌ها شامل برچسب‌گذاری مجدد مجموعه داده‌ها با استفاده از نزدیک‌ترین همسایگان تقریبی (ANN) بر روی نمایش‌های برداری پیام‌ها بود. ANN یک تکنیک است که تطابق‌های تقریبی را برای یک پرسش معین در یک فضای با ابعاد بالا پیدا می‌کند. با استفاده از ANN، هدف تیم بهبود برچسب‌گذاری مجموعه داده بود. برای طبقه‌بندی، تیم UNED از شبکه‌های عصبی استفاده کرد. این شبکه‌های عصبی برای طبقه‌بندی و شناسایی موارد قمار pathological در داده‌های رسانه‌های اجتماعی به کار رفتند.

تیم OBSER-MENH [۱۹]، در هر دو وظیفه اول و دوم چالش شرکت کرد. برای وظیفه اول، رویکرد آن‌ها شامل استفاده از SBERT برای محاسبه نمایش‌های برداری هر انتشار بود. سپس، آن‌ها از توصیف‌های BDI-II به عنوان پرسش‌ها برای رتبه‌بندی علائم افسردگی استفاده کردند. با بهره‌گیری از SBERT و توصیف‌های BDI-II، هدف آن‌ها ارزیابی مرتبط بودن هر علامت در انتشارات بود. در وظیفه دوم، تیم OBSER-MENH بر روی پرداختن به عدم تعادل کلاس و جلوگیری از بیش‌برازش تمرکز کرد. آن‌ها از یک رویکرد تجمعی استفاده کردند که ترکیبی از انواع سه مدل بود. این رویکرد تجمعی به تحلیل وزن‌های جریمه مختلفی که به یک شبکه عصبی پیشرو (FNN) اعمال می‌شد، کمک کرد. این روش به آن‌ها این امکان را داد که چندین مدل را با وزن‌های جریمه متغیر آموزش دهند و به یافتن تعادل بهینه و جلوگیری از بیش‌برازش کمک کنند.

۳-۲-۴ چالش LT-EDI

چالش LT-EDI (Language Technology for Equality, Diversity, Inclusion) به‌عنوان یک وظیفه اشتراکی در زمینه پردازش زبان طبیعی با هدف تشخیص اختلالات روانی از محتوای منتشر شده در شبکه‌های اجتماعی شناخته می‌شود. این چالش بر تشخیص علائم بیماری‌های روحی همچون افسردگی، اضطراب و تمایل به خودکشی متمرکز است. بطور مثال در پژوهش انجام شده توسط آنانتارامان و همکاران [۲]، از مدل BERT

برای دسته‌بندی نشانه‌های افسردگی در محتوای شبکه‌های اجتماعی استفاده کرده‌اند.

وظیفه پنجم در چالش **LT-EDI 2024**: شناسایی استرس در زبان‌های دراویدی

وظیفه پنجم از این چالش به شناسایی استرس در زبان‌های دراویدی، یعنی تامیل و تلگو، پرداخته است. استرس یک حالت احساسی است که ممکن است به دلیل شرایط یا ایده‌هایی که منجر به ناراحتی، خشم یا اضطراب می‌شوند، ایجاد شود. شناسایی زودهنگام استرس و رسیدگی به آن اهمیت زیادی دارد، زیرا در غیر این صورت استرس ممکن است به افسردگی منجر شود. این وظیفه با هدف تشخیص اینکه آیا فردی از طریق پست‌های خود در شبکه‌های اجتماعی استرس دارد یا خیر، ارائه شده است. مدل‌ها باید این پست‌ها را در دو دسته "استرس‌دار" یا "بدون استرس" دسته‌بندی کنند.^۵

مقاله شان و همکاران [۳۰] با موضوع شناسایی استرس در زبان‌های دراویدی، از تکنیک‌های یادگیری ماشین برای تشخیص استرس در محتوای کاربران شبکه‌های اجتماعی به زبان‌های تامیل و تلگو استفاده کرده است. در این مقاله، برای پیش‌پردازش داده‌ها از Label Encoder استفاده شده است که داده‌های متنی را به شکل عددی درآورده و سپس از روش TfIdfVectorizer برای تبدیل متن به ماتریس TF-IDF استفاده کرده‌اند. سه مدل اصلی شامل Naive Bayes، Random Forest و SVM برای طبقه‌بندی داده‌های متنی به کار رفته‌اند. نتایج نشان داد که برای داده‌های تامیل، مدل SVM با دقت ۹۸.۰۹ و F1-Score برابر با ۰.۹۸ بهترین عملکرد را داشت. در داده‌های تلگو، مدل Random Forest با دقت ۹۹.۰۱ و F1-Score برابر با ۰.۹۹ برتر بود. این سیستم به افراد کمک می‌کند تا وضعیت استرس خود را شناسایی کرده و به بهبود سلامت روانی خود بپردازند. در پژوهش اپونن و همکاران [۱]، روش‌هایی برای شناسایی استرس در زبان‌های تامیل و تلگو با استفاده از مدل‌های سنتی یادگیری ماشین بررسی شده است. برای زبان تامیل از مدل Fasttext و برای تلگو از Naive Bayes استفاده شده است. نتایج این مطالعه نشان می‌دهد که مدل Fasttext برای تامیل با F1-score برابر با ۰.۷۷ و مدل Naive Bayes برای تلگو با F1-score برابر با ۰.۷۲ عملکرد قابل توجهی داشته‌اند.

^۵<https://sites.google.com/view/lt-edi-2024/tasks>

۳-۲-۵ نتیجه گیری

باتوجه به پژوهش‌ها و مقاله‌های بررسی شده در این حوزه، می‌توان گفت که در سال‌های اخیر با وجود پیشرفت‌های چشم‌گیر در حوزه توسعه انواع مبدل‌ها، مدل‌های قدیمی‌تر یادگیری ماشین که بر پایه استخراج و یادگیری ویژگی کار می‌کنند، همچنان مورد استفاده پژوهش‌گران بسیاری می‌باشند و توانسته‌اند در بسیاری از مسئله‌ها به نتایج بسیار خوبی دست پیدا کنند.

فصل ۴

آزمایش‌های انجام شده

۴-۱ مقدمه

در این فصل به بررسی کدهای پیاده‌سازی شده، الگوریتم‌ها و مجموعه دادگان مورد استفاده می‌پردازیم و الگوریتم‌ها و خروجی کدها را با یکدیگر با استفاده از معیارهای گوناگون، نمودارها و جداول مقایسه می‌کنیم. مجموعه دادگان مورد استفاده به ترتیب **Sentiment140**، **Depression** و مجموعه دادگان چندکلاسه پژوهش ویلاپرز و همکاران [۳۴] می‌باشند. دو مجموعه داده اول در سامانه **Kaggle** موجود می‌باشند. ولی در راستای استفاده از سومین مجموعه داده، به نویسنده اول مجموعه یک درخواست در قالب ایمیل فرستاده شد تا مجموعه داده را برای ما ارسال کنند. مسئله در مجموعه داده اول و دوم بصورت دودویی می‌باشد و داده‌ها و توییت‌ها تنها به زبان انگلیسی هستند. ولی در مجموعه داده سوم، مسئله از نوع چندکلاسه خواهد بود و ۱۱ کلاس منحصر به فرد وجود دارد. همچنین جمع‌آوردگان این مجموعه داده علاوه بر توییت‌های انگلیسی، توییت‌ها به زبان اسپانیایی را نیز مورد بررسی قرار داده‌اند. و پس از جمع‌آوری و برچسب‌گذاری دادگان از الگوریتم‌های بر پایه استخراج ویژگی یادگیری ماشین استفاده کرده‌اند. البته در ادامه پژوهش دریافتیم که با توجه به محدودیت منابع سخت‌افزاری امکان اجرای مدل‌های زبانی بزرگ بر روی این مجموعه داده را نداریم. ولیکن تحلیل‌های انجام شده بررسی می‌گردند.

۴-۲ مجموعه داده Sentiment140

این مجموعه داده در شاخه مجموعه دادگان تحلیل احساسات کاربران سامانه تویتر قرار می‌گیرد و دارای دو برچسب ۰ و ۱ می‌باشد. برچسب ۰ نشان‌دهنده وجود احساسات منفی در کاربران است و برچسب ۱ بیانگر احساسات مثبت می‌باشد. چند نمونه از داده‌ها را در جدول ۴-۱ می‌توانید مشاهده کنید. در ابتدا تمامی

جدول ۴-۱: پنج ردیف اول مجموعه داده Sentiment140

برچسب (target)	شناسه	متن توییت (TweetText)
0	1467810672	is upset that he can't update his Facebook by ...
0	1467810917	@Kenichan I dived many times for the ball. Man...
0	1467811184	my whole body feels itchy and like its on fire
0	1467811193	@nationwideclass no, it's not behaving at all....
0	1467811372	@Kwesidei not the whole crew

ستون‌ها بجز ستون target و TweetText را حذف می‌کنیم. با توجه به حجم بسیار بالای اولیه این مجموعه داده (۱۶۰ میلیون توییت)، از هر برچسب ۲۰۰۰۰ توییت را بصورت رندوم استخراج می‌کنیم.

۴-۲-۱ پیش‌پردازش

در این مرحله به ترتیب مراحل زیر را بر روی توییت‌ها اعمال می‌کنیم:

۱. بررسی مقادیر تهی در مجموعه داده

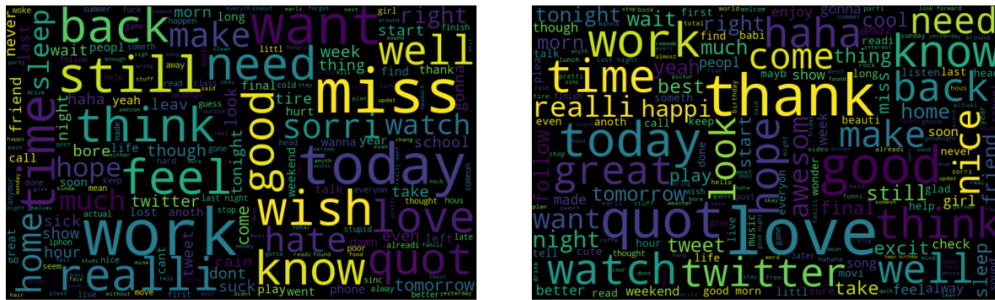
۲. بررسی مقادیر منحصر به فرد برچسب‌ها در مجموعه داده

۳. حذف علائم @ و جایگزینی با رشته‌ای به طول صفر

۴. حذف لینک‌ها و آدرس‌ها

۵. حذف علائم نگارشی، اعداد و علائم خاص

۶. حذف کلمات پایانی (stop words)



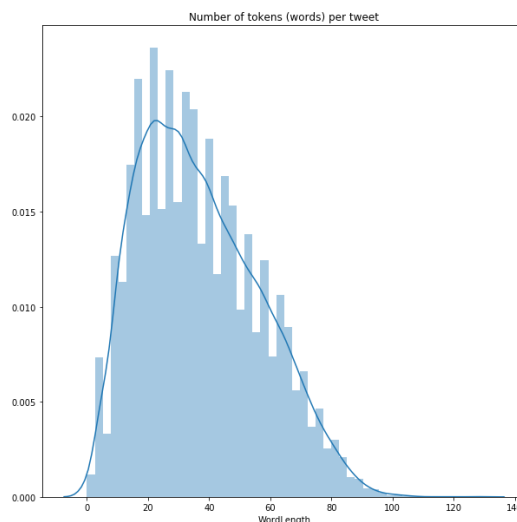
(آ) نمایش ابر کلمات توییت‌ها با برجسب ۱ و محتوای (ب) نمایش ابر کلمات توییت‌ها با برجسب ۰ و محتوای مثبت منفی

شکل ۴-۱: نمایش ابر کلمات توییت‌ها

۷. برگرداندن کلمات به ریشه اولیه (stemming)

۸. حذف کلماتی که کمتر از ۳ لغت دارند

در تصاویر ۴-۱ می‌توانید نمایش ابری کلمات در توییت‌ها با برجسب‌های مجزا را مشاهده کنید. در تصویر ۴-۲ نیز می‌توانید پراکندگی تعداد کلمات توییت‌ها را پس از پیش‌پردازش مشاهده کنید. همان‌طور که مشخص می‌باشد، هیچ توییتی تعداد کلمات بیش از ۱۲۰ ندارد و اکثر توییت‌ها تعداد کلماتی ما بین ۲۰ و ۶۰ دارند.



شکل ۴-۲: پراکندگی تعداد کلمات در توییت‌های استخراج شده پس از پیش‌پردازش

۴-۲-۲ تجزیه داده‌ها به مجموعه آموزشی، تست و اعتبارسنجی

در هنگام بررسی عملکرد الگوریتم‌های کلاسیک مجموعه داده را به نسبت ۸۰/۲۰ به آموزشی و تست تقسیم می‌کنیم. ولی در حالت مدل‌های زبانی بزرگ ۲۰ درصد مجموعه داده را برای اعتبارسنجی نگه می‌داریم. تا در هنگام آموزش مدل در هر دوره بتوانیم ضرر مدل را بر روی داده اعتبارسنجی محاسبه نمائیم.

۴-۲-۳ الگوریتم‌ها و تحلیل خروجی‌ها

در این بخش از چهار الگوریتم XGBoost، Random Forest، Logistic Regression و SVM به عنوان الگوریتم‌های کلاسیک بر روی مجموعه داده اعمال می‌کنیم. همچنین از مدل‌های مبدل زیر نیز به عنوان مدل‌های رمزگذار زبانی استفاده می‌کنیم.

● [google-bert/bert-base-uncased](#) BBUN

● [pig4431/Sentiment140_BERT_5E](#) S140B5E

● [distilbert/distilbert-base-uncased](#) DBBUN [۲۹]

● [austinmw/distilbert-base-uncased-finetuned-tweets-sentiment](#) DBTS

جدول ۴-۲ مقایسه‌ای از عملکرد چهار مدل کلاسیک اعمال شده براساس معیار صحت را نمایش می‌دهد. همان‌طور که مشاهده می‌کنید مدل SVM ضعیف‌ترین عملکرد را با توجه به معیار صحت داشته است. در صورتی که مدل یادگیری گروهی XGBoost بهترین عملکرد را داشته است. دلیل این موضوع می‌تواند این باشد که مدل SVM به درستی ویژگی‌ها را نتوانسته است یاد بگیرد.

جدول ۴-۲: مدل‌های بر پایه استخراج ویژگی بر روی مجموعه داده Sent140 به همراه معیار صحت

مدل	درجه صحت (accuracy)
بردار ماشین پشتیبان (SVM)	۰.۵۸
Logistic Regression	۰.۸۷
Random Forest	۰.۸۶
XGBoost	۰.۹

در آزمایشات انجام گرفته بر روی مدل‌های زبانی در تمامی حالات مقادیر max_length، padding و traunction به ترتیب ۱۵۰، True و True مقداردهی شده‌اند. همچنین در تمامی حالات از تابع AdamW به عنوان optimizer استفاده شده است. مقادیر سایر hyper parameterها را در جدول ۴-۸ به تفکیک مدل‌ها می‌توانید مشاهده کنید. همان‌طور که مشاهده می‌شود، باتوجه به اینکه مدل‌های Distil تعداد پارامتر کمتری دارند و به زمان کمتری برای آموزش نیاز دارند، می‌توانیم از تعداد دور بیشتری برای آموزش مدل استفاده کنیم. البته بدلیل محدودیت‌های سخت‌افزار و کمبود زمان نتوانستیم تعداد دور را بالاتر از این مقادیر ببریم. و یکی از چالش‌های مدل‌های ما در تمام حالات مشکل بیش‌برازش می‌باشد.

جدول ۴-۳: مقادیر برخی hyper parameterها مدل‌های زبانی بر روی مجموعه داده Sent140

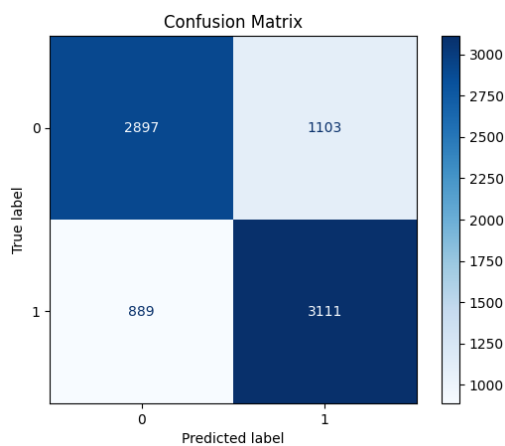
مدل	کاهش وزن	تعداد دور	اندازه دسته آموزشی	اندازه بسته اعتبار سنجی	درجه آموزش
BBUN	۰.۰۱	۵	۱۶	۱۶	2e-5
S140B5E	۰.۰۱	۵	۸	۸	1e-5
DBBUN	۰.۰۱	۱۰	۸	۸	1e-5
DBTS	۰.۰۱	۱۰	۸	۸	1e-5

جدول ۴-۴ نیز مقایسه‌ای از عملکرد چهار مدل زبانی بر اساس معیارها صحت، یادآوری، دقت و امتیاز F1 نمایش می‌دهد. همان‌طور که مشاهده می‌شود، مدل BERT که بر روی مجموعه Sentiment140 از پیش آموزش دیده شده بود، هم اکنون نیز مقدار صحت کل آن بیش از سایر مدل‌ها است. در تصاویر ۴-۳ می‌توانید ماتریس

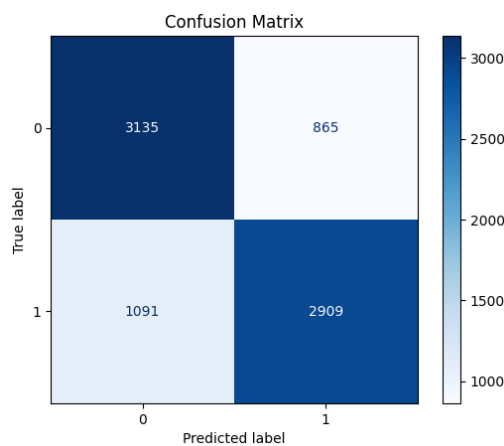
جدول ۴-۴: مقایسه عملکرد LLMها باتوجه به معیارهای صحت، دقت، یادآوری و امتیاز F1

مدل	برچسب	دقت	یادآوری	امتیاز F1	صحت کل
Bert Base Uncased	۰	۰.۷۴	۰.۷۸	۰.۷۶	۰.۷۵
	۱	۰.۷۷	۰.۷۲	۰.۷۴	
Sentiment140 BERT 5E	۰	۰.۷۶	۰.۷۳	۰.۷۵	۰.۷۵۲
	۱	۰.۷۴	۰.۷۷	۰.۷۵	
DistilBert Base Uncased	۰	۰.۷۸	۰.۷	۰.۷۳	۰.۷۴
	۱	۰.۷۲	۰.۸	۰.۷۶	
DistilBert Tweets Sentiment	۰	۰.۷۶	۰.۷۲	۰.۷۴	۰.۷۵۱
	۱	۰.۷۴	۰.۷۷	۰.۷۶	

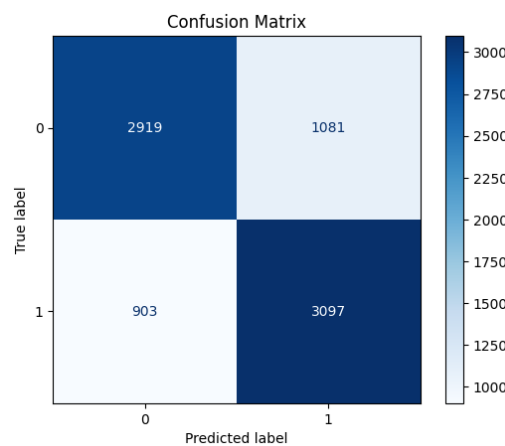
درهم‌ریختگی چهار مدل زبانی اعمال شده را به تفکیک مشاهده کنید. همچنین در تصاویر ۴-۴ نمودارهای



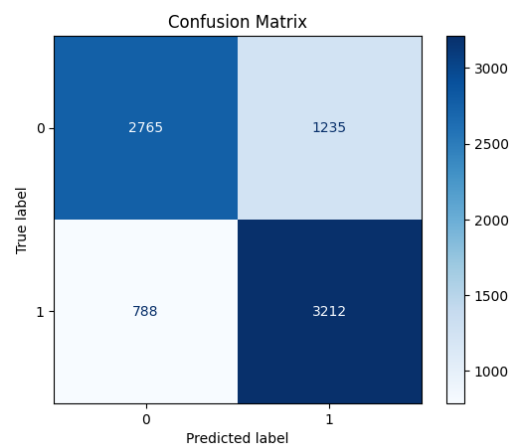
DBTS (ب)



BBUN (آ)



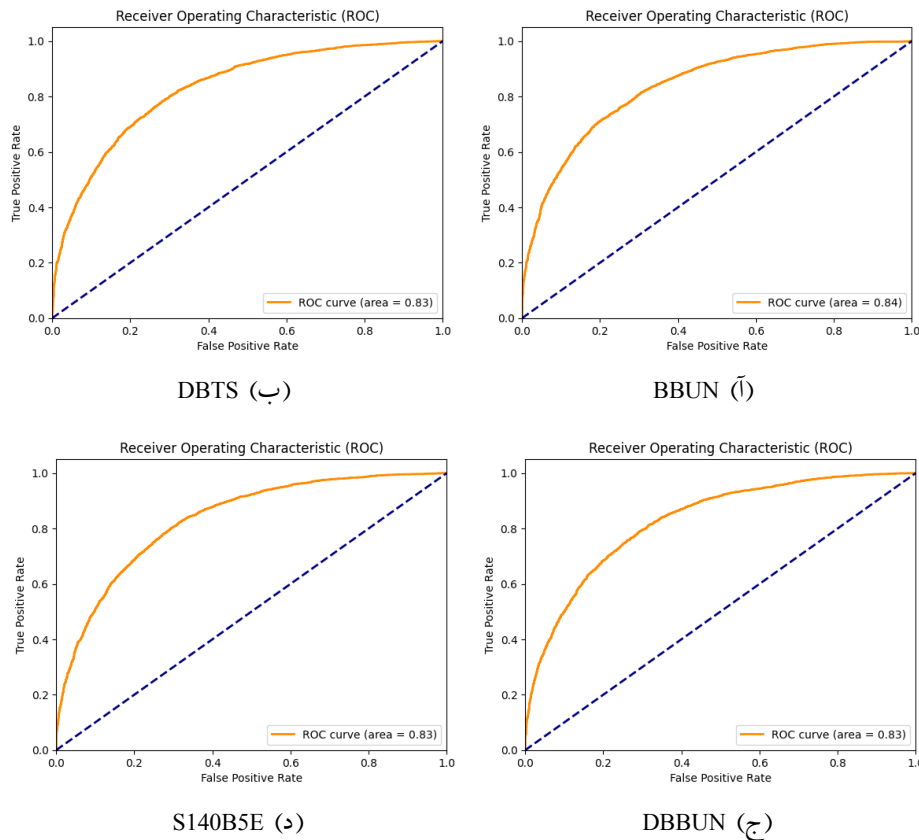
S140B5E (د)



DBBUN (ج)

شکل ۴-۳: نمایش ماتریس‌های در هم‌ریختگی مدل‌های زبانی اعمال شده در آزمایش مجموعه داده Sent

ROC باتوجه به مثبت بود برچسب ۱ آورده شده‌اند. همان‌طور که مشاهده می‌شود در BBUN بهترین مقدار AUC بدست آمده است.



شکل ۴-۴: نمایش نمودارهای ROC مدل‌های زبانی اعمال شده در آزمایش مجموعه داده Sent

۳-۴ مجموعه داده DepressionTweet

این مجموعه داده نیز همانند مجموعه داده بخش پیشین یک مسئله از نوع رده‌بندی دودویی متن کاربران توییتر می‌باشد. جدول ۴-۵ تعدادی از داده‌های این مجموعه را نمایش می‌دهد. البته یکی از تفاوت‌های موجود در این مسئله نیاز به برچسب گذاری داده‌ها می‌باشد. چرا که داده‌های اولیه برچسب‌های دقیقی ندارند و بطور کامل مشخص نشده بود که برچسب ۰ و ۱ بطور دقیق بیان‌گر چه مفاهیمی هستند. در راستای این موضوع از کتابخانه‌های **TextBlob** و **NLTK** استفاده می‌کنیم. در این مرحله در یک حلقه و بصورت متوالی هر توییت را به ورودی تابع `TextBlob(input_tweet).sentiment` می‌دهیم. پس از اعمال این مرحله، ما دو مقدار جدید به نام `polarity` و `subjectivity` به ازای هر توییت خواهیم داشت. نمای جدید مجموعه داده پس از اعمال پیش پردازش و استخراج احساس در جدول ۴-۶ می‌توانید مشاهده کنید.

جدول ۴-۵: پنج ردیف اول مجموعه داده DepressionTweet

شناسه توییت	تاریخ ایجاد	متن توییت
63789..	Sun Aug 30 07:48:37 2015	It's just over 2 years since I was diagnosed w...
63789..	Sun Aug 30 07:31:33 2015	It's Sunday, I need a break, so I'm planning t...
63774..	Sat Aug 29 22:11:07 2015	Awake but tired. I need to sleep but my brain ...
63769..	Sat Aug 29 18:40:49 2015	RT @SewHQ: #Retro bears make perfect gifts and...
63769..	Sat Aug 29 18:40:26 2015	It's hard to say whether packing lists are mak...

۴-۳-۱ پیش‌پردازش

تمامی مراحل مشابه بخش ۴-۲-۱ می‌باشند.

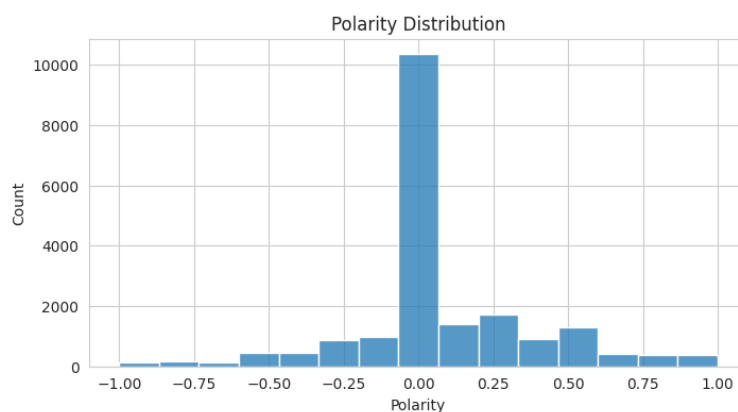
جدول ۴-۶: پنج ردیف اول مجموعه داده DepressionTweet پس از پیش‌پردازش و اعمال تابع TextBlob

subjectivity	polarity	متن توییت
1.000000	0.100000	years since diagnosed anxiety depression today...
0.750000	-0.093750	sunday need break im planning spend little tim...
0.537500	-0.262500	awake tired need sleep brain ideas
0.875000	1.000000	rt sewhq retro bears make perfect gifts great ...
0.370833	-0.045833	hard say whether packing lists making life eas...

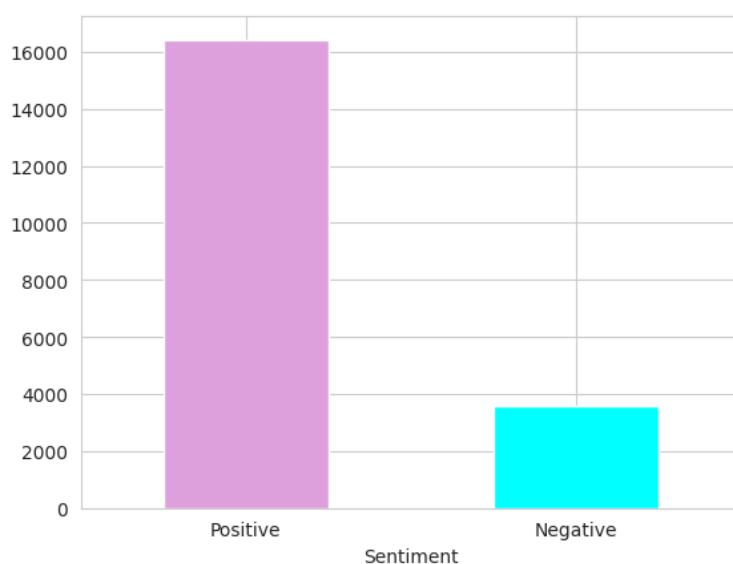
پراکندگی توییت‌ها را با توجه به مقدار خام polarity و بازه‌هایی به طول ۵.۰ می‌توانید در تصویر ۴-۵ مشاهده کنید.

حال در صورتی که مقدار polarity بیش از صفر باشد، برچسب ۱ و در غیر این صورت برچسب ۰ را برای توییت در نظر می‌گیریم. توزیع و پراکندگی توییت‌ها را با توجه به برچسب‌ها می‌توانید در تصویر ۴-۶ مشاهده کنید.

همان‌طور که مشاهده می‌کنید، مجموعه داده بسیار نامتوازن است. در جهت رفع این موضوع نیاز است تا عملیات داده‌افزایی (data augmentation) را انجام دهیم. تا خروجی مدل‌ها گمراه‌کنندگی کمتری داشته باشند. عملیات داده‌افزایی را با استفاده از کتابخانه NLPAug [۱۷] انجام می‌دهیم و با استفاده از حالت SynonymAug داده‌هایی مشابه با داده‌های کلاس Negative ایجاد می‌کنیم و در نهایت داده‌های جدید را به مجموعه داده اولیه اضافه می‌کنیم.



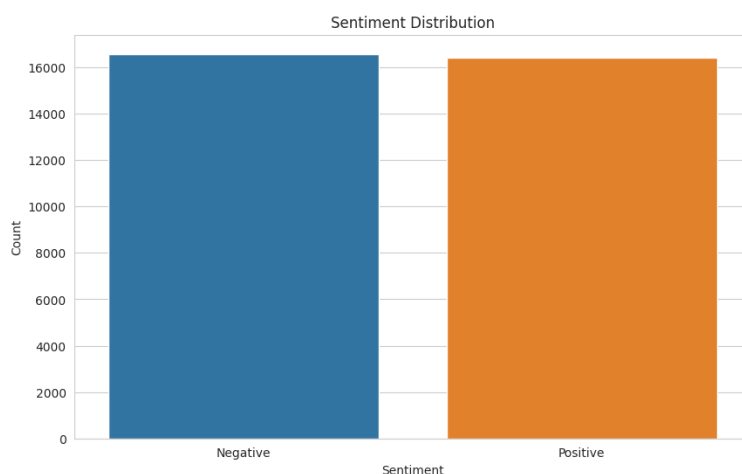
شکل ۴-۵: توزیع داده‌ها براساس مقدار اولیه polarity در مجموعه داده DepressionTweet



شکل ۴-۶: توزیع داده‌ها براساس برجسب در مجموعه داده DepressionTweet

پس از انجام این مرحله و گرفتن خروجی از تعداد توییت‌ها به تفکیک هر برجسب، خروجی زیر چاپ می‌شود. همچنین نمودار توزیع داده‌ها در تصویر ۴-۷ قابل مشاهده است.

```
Sentiment
Negative 16574
Positive 16426
Name: count, dtype: int64
```

شکل ۴-۷: توزیع داده‌ها براساس برچسب در مجموعه داده DepressionTweet پس از اعمال داده‌افزایی

۴-۳-۲ تجزیه داده‌ها به مجموعه آموزشی، تست و اعتبارسنجی

تمامی مراحل مشابه بخش ۴-۲-۲ می‌باشند.

۴-۳-۳ الگوریتم‌ها و تحلیل خروجی‌ها

در این بخش از چهار الگوریتم Naive Bayes، K-Nearest Neighbors، Random Forest و Decision Tree به عنوان الگوریتم‌های کلاسیک بر روی مجموعه داده اعمال می‌کنیم. همچنین از مدل‌های مبدل زیر نیز به عنوان مدل‌های رمزگذار زبانی استفاده می‌کنیم.

● [google-bert/bert-base-uncased](#) BBUN

● [nateraw/bert-base-uncased-emotion](#) BBUNE

● [finiteautomata/bertweet-base-sentiment-analysis](#) BTBSA [۲۶]

● [bhadresh-savani/distilbert-base-uncased-emotion](#) DBBUNE

جدول ۴-۷ مقایسه‌ای از عملکرد چهار مدل کلاسیک اعمال شده براساس معیارهای صحت، دقت، یادآوری و امتیاز F1 را نمایش می‌دهد. همان‌طور که مشاهده می‌کنید مدل‌های KNN و NB ضعیف‌ترین

عملکرد را با توجه به معیار صحت داشته‌اند. در صورتی که مدل یادگیری گروهی Random Forest بهترین عملکرد را داشته است.

جدول ۴-۷: مقایسه عملکرد مدل‌های کلاسیک با توجه به معیارهای صحت، دقت، یادآوری و امتیاز F1

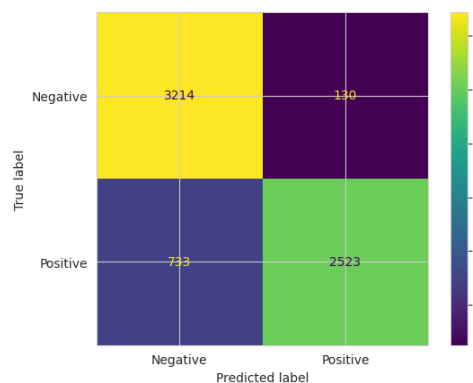
مدل	برچسب	دقت	یادآوری	امتیاز F1	صحت کل
K-Nearest Neighbors	۰	۰.۹۲	۰.۷۹	۰.۸۵	۰.۷۵
	۱	۰.۸۱	۰.۹۳	۰.۸۷	
Naive Bayes	۰	۰.۸۱	۰.۹۶	۰.۸۸	۰.۸۶
	۱	۰.۹۵	۰.۷۷	۰.۸۵	
Random Forest	۰	۰.۹۴	۰.۹۴	۰.۹۴	۰.۹۳
	۱	۰.۹۴	۰.۹۳	۰.۹۴	
Decision Tree	۰	۰.۹۱	۰.۹۳	۰.۹۲	۰.۹۱۶
	۱	۰.۹۳	۰.۹۰	۰.۹۱	

در تصاویر ۴-۸ می‌توانید ماتریس درهم‌ریختگی چهار مدل کلاسیک بر پایه یادگیری و استخراج ویژگی اعمال شده را به تفکیک مشاهده کنید.

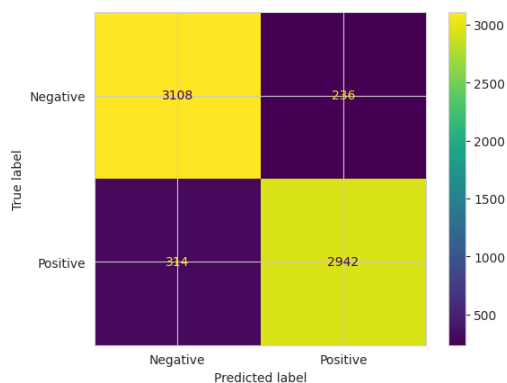
در تصویر ۴-۹ نیز می‌توانید پراکندگی تعداد کلمات توییت‌ها را پس از پیش‌پردازش مشاهده کنید. همان‌طور که مشخص می‌باشد، هیچ توییتی تعداد کلمات بیش از ۱۷۵ ندارد و اکثر توییت‌ها تعداد کلماتی ما بین ۲۵ و ۱۰۰ دارند.

در آزمایشات انجام گرفته بر روی مدل‌های زبانی در تمامی حالات مقادیر max_length، padding و traunction به ترتیب ۱۷۵، True و True مقداردهی شده‌اند. همچنین در تمامی حالات از تابع AdamW به عنوان optimizer استفاده شده است. مقادیر سایر hyper parameterها را در جدول ۴-۸ به تفکیک مدل‌ها می‌توانید مشاهده کنید. همان‌طور که مشاهده می‌شود کاملاً مشابه مجموعه داده بخش قبلی، با توجه به اینکه مدل‌های Distil تعداد پارامتر کمتری دارند و به زمان کمتری برای آموزش نیاز دارند، می‌توانیم از تعداد دور بیشتری برای آموزش مدل استفاده کنیم. البته بدلیل محدودیت‌های سخت‌افزار و کمبود زمان نتوانستیم تعداد دور را بالاتر از این مقادیر ببریم. و یکی از چالش‌های مدل‌های ما در تمام حالات مشکل بیش‌برازش می‌باشد.

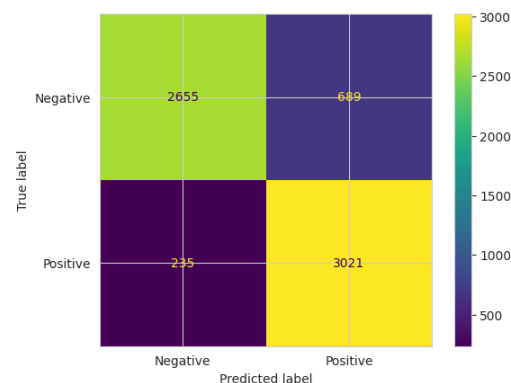
جدول ۴-۹ نیز مقایسه‌ای از عملکرد چهار مدل زبانی بر اساس معیارها صحت، یادآوری، دقت و امتیاز F1 نمایش می‌دهد. همان‌طور که مشاهده می‌شود، مدل BBUNE بهترین عملکرد را نسبت به سایر مدل‌ها



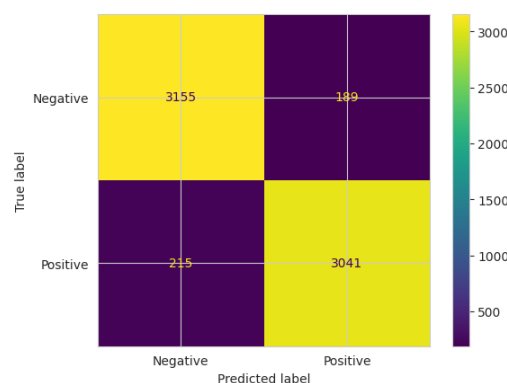
Naive Bayes (ب)



Decision Tree (آ)



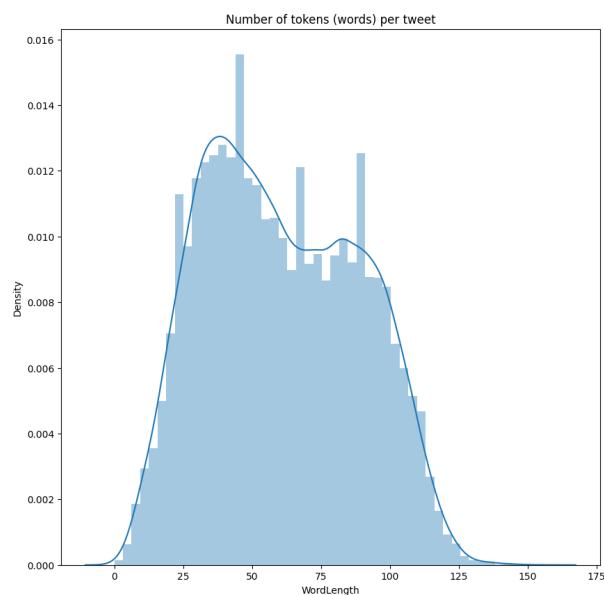
K-Nearest Neighbors (د)



Random Forest (ج)

شکل ۴-۸: نمایش ماتریس‌های در هم‌ریختگی مدل‌های کلاسیک اعمال شده در آزمایش مجموعه داده DepressionTweet

باتوجه به معیار صحت در مجموع داشته است. در این بخش بجز اولین مدل، باقی مدل‌ها بر روی وظایف مشابه وظیفه فعلی از پیش آموزش دیده‌اند. همچنین در تصاویر ۴-۱۰ نمودارهای ROC باتوجه به مثبت بود برچسب ۱ آورده شده‌اند. همان‌طور که مشاهده می‌شود هر سه مدل مقدار AUC برابر با ۹۷.۰ بدست آورده‌اند.



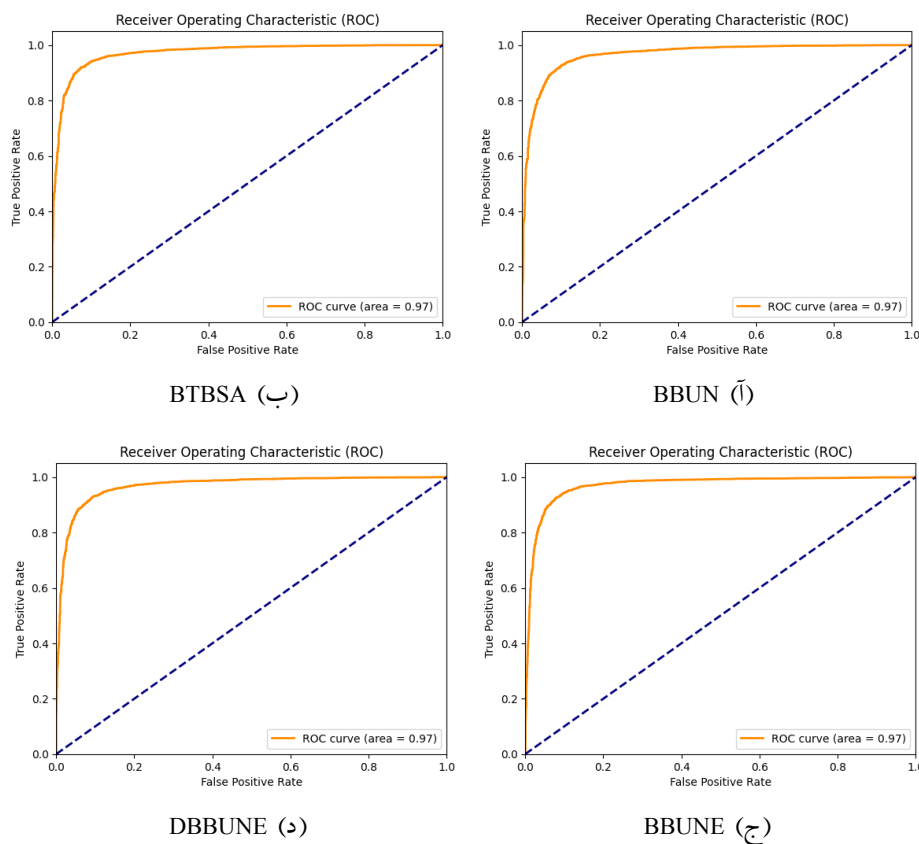
شکل ۴-۹: پراکندگی تعداد کلمات در توییت‌های استخراج شده پس از پیش‌پردازش و داده‌افزایی

جدول ۴-۸: مقادیر برخی hyper parameter مدل‌های زبانی بر روی مجموعه داده Depression Tweet

مدل	کاهش وزن	تعداد دور	اندازه دسته آموزشی	اندازه بسته اعتبار سنجی	درجه آموزش
BBUN	۰.۰۱	۵	۸	۸	2e-5
BTBSA	۰.۰۱	۵	۸	۸	2e-5
BBUNA	۰.۰۱	۵	۸	۸	2e-5
DBBUNE	۰.۰۱	۱۰	۸	۸	2e-5

جدول ۴-۹: مقایسه عملکرد LLMها با توجه به معیارهای صحت، دقت، یادآوری و امتیاز F1

مدل	برچسب	دقت	یادآوری	امتیاز F1	صحت کل
Bert Base Uncased	۰	۰.۸۸	۰.۹۳۶	۰.۹۱	۰.۹۰
	۱	۰.۹۳	۰.۸۷	۰.۹۰	
Bert Tweet Base Sentiment Analysis	۰	۰.۹۲	۰.۹۳	۰.۹۲	۰.۹۲
	۱	۰.۹۲	۰.۹۲	۰.۹۲	
Bert Base Uncased Emotion	۰	۰.۹۴	۰.۹۱	۰.۹۲	۰.۹۳
	۱	۰.۹۱	۰.۹۴	۰.۹۳	
DistilBert Uncased Emotion	۰	۰.۹۱	۰.۹۲	۰.۹۲	۰.۹۲
	۱	۰.۹۲	۰.۹۱	۰.۹۱	



شکل ۴-۱۰: نمایش نمودارهای ROC مدل‌های زبانی اعمال شده در آزمایش مجموعه داده Depression Tweet

۴-۴ مجموعه داده ویلاپرز و همکاران

این مجموعه داده بر خلاف دو مجموعه داده پیشین از نوع چندکلاسه (۱۱ برچسب) می‌باشد. همچنین به دلیل محرمانه بودن اطلاعات، به جمع‌آوردندگان مجموعه ایمیلی با مضمون درخواست برای دریافت دادگان فرستاده شد. این مجموعه داده که بیش از یک گیگابایت حجم دارد، دارای چندین پوشه و بخش‌بندی می‌باشد و در اولین قدم نیاز است که تمامی داده‌ها را به یک دیگر متصل کنیم. همچنین این مجموعه داده متشکل از توییت‌ها به زبان انگلیسی و فارسی می‌باشد [۳۴]. در راستای این موضوع توییت‌های انگلیسی و اسپانیایی را در دو فایل csv مجزا به یک دیگر متصل کردیم. مجموعه توییت‌های اسپانیایی شامل ۲۶۶۳۷۶۱ توییت و مجموعه توییت‌های انگلیسی شامل ۵۱۷۰۴۹۳ توییت می‌باشد.

ویلاپرز و همکاران در مقاله خود [۳۳] چندین روش مبتنی بر استخراج ویژگی و همچنین چند مدل برپایه

شبکه‌های عصبی پیچشی بر روی دادگان خود اعمال کرده‌اند. هدف ما بررسی چند مدل زبانی بزرگ بر روی این مجموعه بود. اما متأسفانه بدلیل حجم بالا مجموعه دادگان، حتی بعد از چندین بار نمونه‌گیری و پیش‌پردازش به نتایج خوبی دست نیافتیم و به همین دلیل در این پژوهش به گزارشی از تحلیل ابتدایی این مجموعه داده اکتفا می‌کنیم.

جدول ۴-۱۰ نمایی از چند داده اول این مجموعه را نمایش می‌دهد (بدلیل کمبود عرض صفحه نتوانستیم تمام ویژگی‌ها و ابعاد مجموعه داده را نمایش دهیم).

جدول ۴-۱۰: پنج ردیف اول مجموعه داده ویلاپرز و همکاران

کلاس	شناسه توییت	روز	زمان	توییت
ASD	14383..	2021-09-16	02:27:56	Ya sÃ¡quenme ...
ASD	14374..	2021-09-13	18:27:23	@USUARIO AsÃ...
ASD	14372..	2021-09-13	03:24:29	@USUARIO Cuanta..
ASD	14372..	2021-09-13	03:16:48	@USUARIO Tienes ex bonitas..
ASD	14361..	2021-09-10	01:44:54	ESTO HTTPURL

۴-۴-۱ پیش‌پردازش

در این مرحله به ترتیب مراحل زیر را بر روی توییت‌ها اعمال می‌کنیم و همچنین از کدهای موجود در پروژه [Github](#) ویلاپرز و همکاران استفاده می‌کنیم. یکی از توابع استفاده شده در این پیش‌پردازش preprocess_tweet از کتابخانه [pysentimiento](#) می‌باشد. و همچنین از کتابخانه [spaCy](#) نیز استفاده شده است.

۱. حذف مقادیر تکراری براساس ویژگی tweet_id

۲. بررسی مقادیر تهی در مجموعه داده

۳. بررسی مقادیر منحصر به فرد برچسب‌ها در مجموعه داده و حذف داده‌ها با برچسب‌های پرت و مشکل دار

۴. حذف توییت‌ها با طول بالاتر از مقدار نرمال (این مرحله با استفاده از روابط چارک‌ها و IQR انجام شده است. در هر دو حالت توییت‌های انگلیسی به دلیل تعداد کم توییت‌ها با طول بیش از ۳۰۰، این توییت‌ها حذف شده‌اند.)

۵. حذف لینک‌ها و آدرس‌ها

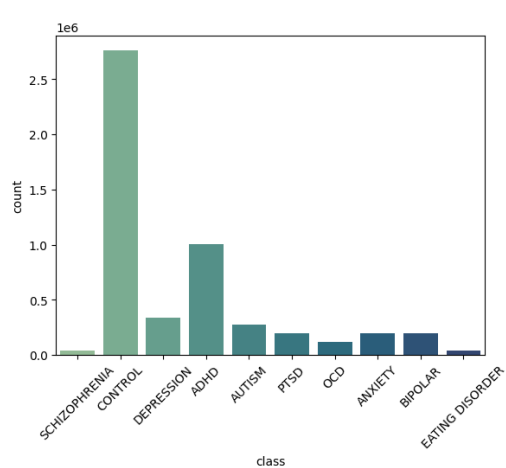
۶. حذف علائم نگارشی، اعداد و علائم خاص

۷. حذف کلمات پایانی (stop words)

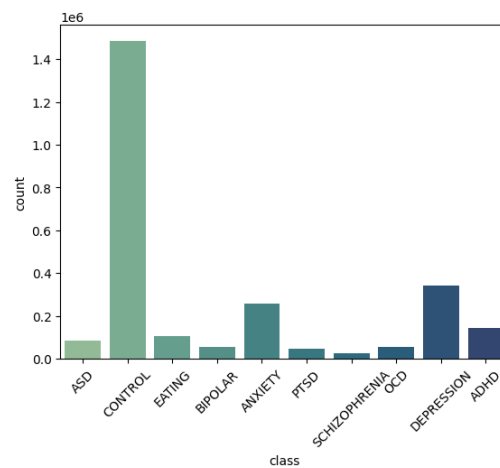
۸. کوچک‌سازی همه حروف

۹. حذف تمامی هشتک‌ها

توزیع توییت‌ها براساس برچسب‌ها را می‌توانید در تصویر ۴-۱۱ مشاهده کنید.



(ب) توزیع توییت‌های انگلیسی



(آ) توزیع توییت‌های اسپانیایی

شکل ۴-۱۱: توزیع توییت‌های انگلیسی و اسپانیایی با توجه به برچسب‌ها پس از حذف برچسب‌های مشکل‌دار

فصل ۵

نتیجه گیری و پیشنهادها

۵-۱ نتیجه گیری و جمع بندی

در این پایان نامه، به تشخیص بیماری های روحی و روان شناختی از جمله افسردگی، استرس، اعتیاد و تمایل به خودکشی از طریق تحلیل محتوای متنی کاربران در شبکه های اجتماعی پرداخته شد. با توجه به رشد فزاینده استفاده از این پلتفرم ها، تحلیل داده های کاربران می تواند به شناسایی سریع تر مشکلات روان شناختی کمک کند. هدف اصلی این پروژه، توسعه الگوریتم های یادگیری ماشین و مدل های زبانی پیشرفته برای تحلیل و دسته بندی متون کاربران بود. در این راستا، الگوریتم های مختلفی همچون SVM، XGBoost، درخت تصمیم و جنگل تصادفی به همراه مدل های جدیدتر مانند رمزگذارهای خانواده BERT مورد استفاده قرار گرفتند تا عملکرد آن ها در تشخیص این اختلالات مورد ارزیابی قرار گیرد.

نتایج نشان داد که مدل های زبانی بزرگ و مبدل ها به دلیل قابلیت های پیشرفته در تحلیل معنایی متن، دقت بیشتری در تشخیص مشکلات روان شناختی دارند. با این حال، محدودیت های سخت افزاری مانع از آزمایش کامل برخی مدل های پیچیده شد که این موضوع می تواند به عنوان یک زمینه برای تحقیقات آینده در نظر گرفته شود. در مجموع، این پروژه گامی مؤثر در جهت استفاده از تکنولوژی های پیشرفته برای بهبود سلامت روانی افراد از طریق شبکه های اجتماعی است و می تواند به توسعه سیستم های تشخیصی خودکار در این حوزه کمک کند.

۵-۲. پیشنهادها

باتوجه به اینکه چالش بسیاری از دانشجویان گرایش هوش مصنوعی نبود منابع سخت‌افزاری پیشرفته به تعداد کافی در محیط دانشگاه می‌باشد، دانشگاه و دانشکده می‌توانند برای این زمینه بودجه کافی و مناسب اختصاص بدهند تا دانشجویان بتوانند پژوهش‌های خود را با کیفیت بالاتری انجام بدهند.

۵-۳. کارهای آینده

قطعا ارزش یک کار پژوهشی به تعداد دفعات آزمایش‌ها و تکرار آن‌ها می‌باشد. در این راستا اگر بتوانیم با استفاده از الگوریتم‌هایی همانند T-test و P-value خروجی مدل‌های خود را بیازماییم، می‌توانیم از درستی خروجی و معیارها اطمینان بیشتری داشته باشیم.

همچنین هدف داریم با استفاده از یک مدل تعمیم‌دهنده (adapter) مدل‌های آموزش دیده شده را به یک مدل تبدیل نمائیم تا در صورت استفاده در پروژه‌های بزرگ‌تر، کدهای تکراری کم‌تر داشته باشیم. باتوجه به پیشرفت‌های روزانه در پردازش زبان طبیعی و انتشار مدل‌های گوناگون با توانایی‌ها و دقت‌های بالاتر، تلاش داریم تا بتوانیم از آن‌ها نیز در رده‌بندی متن استفاده کنیم. همچنین استفاده از روش‌های نوین و به‌روزتر همانند Few Shot Learning و Zero Shot Learning مورد توجه ما می‌باشد.

کتاب نامه

- [1] Alex Eponon, A., Batyrshin, I., and Sidorov, G. Pinealai_StressIdent_LT-EDI@EACL2024: Minimal configurations for stress identification in Tamil and Telugu. in *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion* (St. Julian's, Malta, Mar. 2024), B. R. Chakravarthi, B. B. P. Buitelaar, T. Durairaj, G. Kovács, and M. Á. García Cumberras, eds. , Association for Computational Linguistics, pp. 152–156.
- [2] Anantharaman, K., Angel, S., Sivanaiah, R., Madhavan, S., and Rajendram, S. M. Ssn_mlr1@ lt-edi-acl2022: Multi-class classification using bert models for detecting depression signs from social media text. in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (2022), pp. 296–300.
- [3] Bezerra, E., Santos, L. d., Nascimento, R. F., Lopes, R. P., and Guedes, G. P. Nailp at erisk 2023: search for symptoms of depression. in *24th Working Notes of the Conference and Labs of the Evaluation Forum (CLEF-WN)* (2023), volume 3497, CEUR-WS, pp. 639–661.
- [4] Bucur, A.-M. Utilizing chatgpt generated data to retrieve depression symptoms from social media. *arXiv preprint arXiv:2307.02313* (2023).
- [5] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805* (2018).
- [6] Fabregat, H., Duque, A., Araujo, L., and Martínez-Romo, J. Nlp-uned-2 at erisk 2023: Detecting pathological gambling in social media through dataset relabeling and neural networks. in *CLEF (Working Notes)* (2023), pp. 672–683.
- [7] Fields, J., Chovanec, K., and Madiraju, P. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access* 12 (2024), 6518–6531.

- [8] Gasparetto, A., Marcuzzo, M., Zangari, A., and Albarelli, A. A survey on text classification algorithms: From text to predictions. *Information* 13, 2 (2022), 83.
- [9] Grigore, D.-N., and Pintilie, I. Transformer-based topic modeling to measure the severity of eating disorder symptoms. in *CLEF (Working Notes)* (2023), pp. 684–692.
- [10] Ilias, L., Mouzakitis, S., and Askounis, D. Calibration of transformer-based models for identifying stress and depression in social media. *IEEE Transactions on Computational Social Systems* 11, 2 (2023), 1979–1990.
- [11] Kim, J., Lee, J., Park, E., and Han, J. A deep learning model for detecting mental illness from user content on social media. *Scientific reports* 10, 1 (2020), 11846.
- [12] Kumar, P., Samanta, P., Dutta, S., Chatterjee, M., and Sarkar, D. Feature based depression detection from twitter data using machine learning techniques. *Journal of Scientific Research* 66, 2 (2022), 220–228.
- [13] Larrayoz, X., Lebeña, N., Casillas, A., and Pérez, A. Representation exploration and deep learning applied to the early detection of pathological gambling risks. in *CLEF (Working Notes)* (2023), pp. 693–705.
- [14] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364* (2020).
- [15] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 2 (2022), 1–41.
- [16] Lin, T., Wang, Y., Liu, X., and Qiu, X. A survey of transformers. *AI Open* 3 (2022), 111–132.
- [17] Ma, E. Nlp augmentation. <https://github.com/makcedward/nlpaug>, 2019.
- [18] Mármol-Romero, A. M., del Arco, F. M. P., and Montejo-Ráez, A. Sinai at erisk@ clef 2023: Approaching early detection of gambling with natural language processing. in *CLEF (Working Notes)* (2023), pp. 743–751.
- [19] Martínez-Romo, J., Araujo, L., Larrayoz, X., Oronoz, M., and Pérez, A. Obser-menh at erisk 2023: Deep learning-based approaches for symptom detection in depression and early identification of pathological gambling indicators. in *CLEF (Working Notes)* (2023), pp. 706–717.

- [20] Maupomé, D., Soulas, T., Rancourt, F., Cantin-Savoie, G., Winterstein, G., Mosser, S., and Meurs, M.-J. Lightweight methods for early risk detection. in *CLEF (Working Notes)* (2023), pp. 718–726.
- [21] Merhbene, G., Puttick, A. R., and Kurpicz-Briki, M. Bfh-ami at erisk@ clef 2023. in *CLEF (Working Notes)* (2023).
- [22] Molina, A., Huang, X., Hurtado, L.-F., and Pla, F. Elirf-upv at erisk 2023: Early detection of pathological gambling using svm. in *CLEF (Working Notes)* (2023), pp. 736–742.
- [23] Orabi, A. H., Buddhitha, P., Orabi, M. H., and Inkpen, D. Deep learning for depression detection of twitter users. in *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic* (2018), pp. 88–97.
- [24] Pan, R., García-Díaz, J. A., and Valencia-García, R. Umuteam at erisk@ clef 2023 shared task: Transformer models for early detection of pathological gambling, depression, and eating disorder. in *CLEF (Working Notes)* (2023), pp. 752–762.
- [25] Patwardhan, N., Marrone, S., and Sansone, C. Transformers in the real world: A survey on nlp applications. *Information 14*, 4 (2023), 242.
- [26] Pérez, J. M., Giudici, J. C., and Luque, F. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks, 2021.
- [27] Recharla, N., Bolimera, P., Gupta, Y., and Madasamy, A. K. Exploring depression symptoms through similarity methods in social media posts. in *CLEF (Working Notes)* (2023), pp. 763–772.
- [28] Sakib, F. A., Choudhury, A. A., and Uzuner, O. Mason-nlp at erisk 2023: deep learning-based detection of depression symptoms from social media texts. *arXiv preprint arXiv:2310.10941* (2023).
- [29] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv abs/1910.01108* (2019).
- [30] Shanmugavadivel, K., Subramanian, M., J, M., S, M., and B, R. KEC_AI_MIRACLE_MAKERS@LT-EDI-2024: Stress identification in Dravidian languages using machine learning techniques. in *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion* (St. Julian’s, Malta, Mar. 2024),

- B. R. Chakravarthi, B. B. P. Buitelaar, T. Durairaj, G. Kovács, and M. Á. García Cumberas, eds., Association for Computational Linguistics, pp. 277–281.
- [31] Talha, A., and Basu, T. A natural language processing based risk prediction framework for pathological gambling. in *CLEF (Working Notes)* (2023), pp. 782–790.
- [32] Thompson, H., Cagnina, L., and Errecalde, M. Strategies to harness the transformers’ potential: Unsl at erisk 2023. *arXiv preprint arXiv:2310.19970* (2023).
- [33] Villa-Pérez, M. E., Trejo, L. A., Moin, M. B., and Stroulia, E. Extracting mental health indicators from english and spanish social media: A machine learning approach. *IEEE Access* 11 (2023), 128135–128152.
- [34] Villa-Pérez, M. E., and Trejo, L. A. Twitter dataset for mental disorders detection, feb 2023.
- [35] Wang, Y., and Inkpen, D. uottawa at erisk 2023: Search for symptoms of depression. in *CLEF (Working Notes)* (2023), pp. 805–812.