مطالعه‌ای روی تحلیل احساسات عربی توییتر برای تشخیص افسردگی بر اساس تجمیع اطلاعات زبانی، به‌کمک تکنیک‌های یادگیری ماشین

علیرضا محمودی‌فرد [۱] ، علی ملکی [۲*]

[۱] دانشجوی‌کارشناسی‌ارشد ناپیوسته مدیریت صنعتی، دانشکده علوم انسانی، دانشگاه شاهد (و فارغ‌التحصیل کارشناسی ارشد مهندسی برق و مدرس دانشگاه-ها)، تهران، ایران، alireza10.m10@gmail.com

[۲] دانشجوی کارشناسی‌ارشد ناپیوسته مهندسی مخابرات دانشگاه تربیت دبیر شهید رجایی، تهران، ایران، A.malekibme@gmail.com

اطلاعات مقاله

تاریخچه مقاله:
تاریخ دریافت مقاله: ۱۴۰۱/۱۲/۱۰
تاریخ پذیرش مقاله: ۱۴۰۲/۰۱/۱۵
تاریخ انتشار مقاله: ۱۴۰۲/۰۱/۱۷

کلمات کلیدی:
داده‌کاوی
افسردگی
فراگیری ماشین
تحلیل احساسات
شبکه اجتماعی توییتر
Machine learning

چکیده

پلتفرم‌های رسانه‌های اجتماعی امروزه به‌طور فزاینده‌ای مورد استفاده قرار می‌گیرند؛ زیرا بسیاری از مردم در سراسر جهان در تعامل، ارتباط و اشتراک‌گذاری محتوا با دیگران هستند؛ کاربران شبکه‌های اجتماعی اغلب احساسات و عواطف خود را در پست‌های خود آشکار می‌کنند؛ در زمینه تصمیم‌گیری کیفی، متغیرهای زبانی کاربران این شبکه‌ها، اغلب اطلاعاتی را بیان می‌کنند که به‌جای کمیت دقیق، مربوط به رتبه‌بندی ترتیبی است و بنابراین چالش در چگونگی تعیین وزن‌های منطبق با واقعیت و معنا بخشیدن به تصمیم آن‌ها است. رسانه‌های اجتماعی به یک منبع آنلاین حیاتی برای مطالعه زبان مورد استفاده کاربران رسانه‌های اجتماعی، برای بیان مسائل مربوط به سلامت روان خود تبدیل شده‌اند که می‌توانند به شناسایی افراد در معرض آسیب کمک کنند؛ اکنون محققان بیشتر به سلامت روان از طریق رسانه‌های اجتماعی علاقمند شده‌اند؛ شبکه اجتماعی توییتر، با موفقیت برای بررسی چندین بیماری روانی از جمله اضطراب، افسردگی، افکار آسیب رساندن به خود و خودکشی پیاده‌سازی شده است؛ افسردگی عامل اصلی بیماری و ناتوانی در سراسر جهان است و تعداد افراد مبتلا به اختلالات روانی رایج در سطح جهان در حال افزایش است. در این مقاله، مدلی مورد مطالعه و تحلیل قرار گرفته است که می‌تواند توییت‌های عربی را بر اساس ویژگی‌های افسردگی انتخاب شده توسط متخصصان سلامت طبقه‌بندی کند؛ در جمع‌آوری داده‌ها، توییت‌ها از API توییتر جمع‌آوری شده‌اند؛ سپس، تکنیک‌های یادگیری ماشینی تحت نظارت، برای استخراج توییت‌هایی با بیشترین ویژگی‌های افسردگی اعمال شده است؛ پس از آن، دقت در بین الگوریتم‌های یادگیری ماشینی تحت نظارت اعمال شده، ارزیابی شده است تا بهترین الگوریتم برای مدل خود شناسایی شود. نویسندگان این مقاله و محققان مقالات مرجع، بر این باورند که این پروژه می‌تواند توسط پزشکان بهداشت برای کمک به تشخیص و ارائه کمک به کاربران افسرده شبکه اجتماعی توییتر استفاده شود.

۱ - مقدمه

◄ در دهه‌های اخیر، استفاده از پلتفرم‌های رسانه‌های اجتماعی، افزایش یافته است؛ مردم می‌توانند ارتباط برقرار کنند و محتوا را با دیگران به اشتراک بگذارند؛ فیس‌بوک و توییتر، محبوب‌ترین پلتفرم-ها هستند؛ کاربران شبکه‌های اجتماعی، می‌توانند احساسات خود را در پست‌های خویش آشکار کنند [۱ و ۳].

در سال‌های اخیر، مطالعات بیشتر به سلامت روان از طریق پلتفرم-های رسانه‌های اجتماعی علاقه‌مند شده‌اند؛ بسیاری از مطالعات ارتباط زبان و الگوهای استفاده از رسانه‌های اجتماعی را با چندین بیماری روانی از جمله استرس، افسردگی، اضطراب و خودکشی بررسی کرده‌اند؛ کاربران شبکه‌های اجتماعی اغلب احساسات، افکار و نظرات خود را با دیگران بیان می‌کنند؛ محتوای فعالیت‌های کاربران می‌تواند منبع ارزشمندی از اطلاعات باشد که می‌تواند برای شناسایی و تشخیص علائم افسردگی در کاربران رسانه‌های اجتماعی مورد استفاده قرار گیرد [۲۳ و ۲۷]؛ پست‌های رسانه‌های اجتماعی، به ثبت ویژگی‌های رفتاری مرتبط با تفکر، خلق و خو، ارتباطات، نظرات و فعالیت‌های فرد کمک می‌کنند؛ زبان و احساسات مورد استفاده در پست‌های رسانه‌های اجتماعی، می‌تواند به احساس بی‌ارزشی، درماندگی، گناه و نفرت از خود اشاره کند که مشخصه افسردگی اساسی است [۴]. در این مقاله، بررسی مدلی برای شناسایی علائم افسردگی انجام شده است؛ این مدل می‌تواند توییت‌های عربی را بر اساس ویژگی‌های افسردگی طبقه‌بندی کند تا به پزشکان در تصمیم‌گیری کمک کند؛ در جمع آوری داده‌ها، توییت‌ها از API توییتر جمع‌آوری شده است؛ سپس، تکنیک‌های یادگیری ماشینی تحت نظارت، برای استخراج توییت‌هایی با بیشترین ویژگی‌های افسردگی اِعمال شده است؛ پس از آن، دقت در میان الگوریتم‌های یادگیری ماشینی تحت نظارت کاربردی ارزیابی شده است تا بهترین الگوریتم برای مدل خود شناسایی شود [۳ و ۲۷].

۲ - زمینه

◄ یادگیری ماشینی، به‌دلیل افزایش کاربردهای داده‌کاوی، زمینه‌ای سریع در حال رشد است؛ مطالعات و برنامه‌های کاربردی اخیر، بر مشکل یادگیری ماشین تمرکز دارند؛ در یادگیری ماشینی، داده‌ها برای انجام وظایف جمع‌آوری، مرتب‌سازی، جذب و طبقه‌بندی اطلاعات، پردازش می‌شوند [۱۳]؛ تکنیک‌های یادگیری ماشین، به-شدت با داده‌کاوی مرتبط هستند؛ یادگیری ماشینی، نشان می‌دهد که چگونه مدل‌ها می‌توانند عملکرد خود را بر اساس داده‌ها یاد بگیرند یا بهبود بخشند. هدف اصلی یادگیری ماشینی، این است که مدل به‌طور خودکار یاد بگیرد که چگونه الگوهای پیچیده را شناسایی و استخراج کند و بر اساس داده‌ها، تصمیمات هوشمندانه بگیرد؛ تکنیک‌های یادگیری ماشین، به دو دسته یادگیری نظارت شده و یادگیری بدون نظارت [۸] طبقه‌بندی می‌شوند [۲۷].

تجزیه و تحلیل داده‌های رسانه‌های اجتماعی، امکان تشخیص دقیق علائم افسردگی را قبل از اینکه به مراحل شدیدتر افسردگی برسند، می‌دهد؛ این امر، امکان توصیه راهبردهایی را برای پیشگیری و درمان افسردگی در مراحل اولیه فراهم می‌کند؛ سایت‌های رسانه‌های اجتماعی برای شناسایی افسردگی استفاده شده‌اند که توئیتر بیشترین استفاده را دارد [۲۸ و ۲۹]؛ بیش از ۴۳۶ میلیون کاربر فعال ماهانه توییتر بیش از ۵۰۰ میلیون توییت در روز ارسال می‌کنند و توییتر را به نهمین وب‌سایت محبوب اینترنت از نظر محبوبیت تبدیل می‌کند [۳۰]؛ هر کاربر ثبت‌نام شده می‌تواند نظرات خود را در ۱۴۰ کاراکتر در یک زمان منتشر کند؛ توییت‌ها اغلب عمومی می‌شوند و ممکن است با استفاده از API توییتر جمع‌آوری و تجزیه و تحلیل شوند [۳۱]؛ علاوه بر این، API توییتر جستجوهای پیچیده-ای مانند بازیابی هر توییت در مورد یک موضوع خاص را امکان‌پذیر می‌کند؛ تجزیه و تحلیل گفتار و متن از طریق پردازش زبان طبیعی (NLP) انجام می‌شود؛ در طول توسعه فناوری رایانه، رویکردهای زبانی به الگوریتم‌های رایانه‌ای تبدیل شدند؛ پردازش زبان طبیعی در ابتدا برای بررسی وظایف کلاسیک، مانند ساختار گرامر در کتاب‌ها استفاده می‌شد؛ با این حال، به تفسیر دیدگاه‌های انسانی، از جمله ایمیل، محتوای آنلاین، بررسی‌ها، پست‌های شبکه‌های اجتماعی و مقالات رسانه‌ای گسترش یافته است [۳۲ و ۳۳]؛ از بسیاری جهات، NLP می‌تواند زبان‌های مختلفی را پردازش کند [۲۸]؛ به‌عنوان مثال، با استفاده از تحلیل احساسات (SA)، می‌توانید جنبه‌های مثبت و منفی یک متن یا گفتار را شناسایی کنید (بوشان و شارما، ۲۰۲۱)؛ این روش ممکن است برای تجزیه و تحلیل سطوح افسردگی افراد بر اساس پست‌های رسانه‌های اجتماعی و رتبه‌بندی احساسات منفی سازگار شود؛ بسیاری از الگوریتم‌های پیش‌بینی و رویکردهای بهینه‌سازی برای مشاهده الگوها در داده‌ها و ایجاد بینش بر اساس آن مشاهدات، مانند یادگیری عمیق (DL) و یادگیری ماشین (ML) وجود دارد؛ به‌عنوان مثال، زمانی‌که محققان از روش‌های طبقه‌بندی باینری برای تجزیه و تحلیل متن استفاده می‌کنند، از الگوریتم‌های یادگیری ماشین سنتی (ML) مانند جنگل تصادفی (RF)، ماشین‌های بردار پشتیبان (SVM)، درخت‌های تصمیم‌گیری (DT) و غیره استفاده می‌کنند [۳۴- ۳۶].

۳ – الگوریتم‌های نظارت شده

◄ یادگیری تحت نظارت معادل طبقه‌بندی [۸ و ۲۷] است؛ یک مجموعه آموزشی و یک مجموعه تست برای دسته‌بندی داده‌ها استفاده می‌شود؛ ویژگی‌های ورودی و برچسب‌های کلاس مربوط به آن‌ها در مجموعه آموزشی گنجانده شده است؛ مجموعه داده آموزشی برای ساخت مدل طبقه‌بندی استفاده می‌شود، که هدف آن دسته‌بندی ویژگی‌های ورودی به برچسب‌های کلاس منطبق است، در حالی‌که مجموعه داده آزمایشی برای آزمایش اعتبار مدل با پیش‌بینی برچسب‌های کلاس ویژگی‌های دیده نشده استفاده می‌شود؛ برای دسته‌بندی مجموعه‌های داده، الگوریتم‌های یادگیری ماشینی مانند Naive Bayes (NB)، درخت تصمیم (C4.5، ID3 و C5)، و ماشین‌های بردار پشتیبان[۱] استفاده می‌شوند [۵ و ۲۷].

۴- الگوریتم‌های بدون نظارت

◄ خوشه‌بندی معادل یادگیری بدون نظارت است؛ از آنجا که مجموعه داده‌های ورودی دارای برچسب کلاس نیستند، فرآیند یادگیری بدون نظارت است؛ به‌طور معمول، خوشه‌بندی می‌تواند برای شناسایی کلاس‌های درون داده‌ها استفاده شود؛ تکنیک خوشه‌بندی گروهی از اشیاء را بر اساس شباهت خواص آن‌ها به خوشه‌ها دسته-بندی می‌کند [۷]؛ در یک خوشه، اشیاء شبیه به یکدیگر هستند و اشیاء از خوشه‌های مختلف غیر مشابه هستند؛ شباهت‌ها و عدم تشابه‌ها با استفاده از مقادیر مشخصه‌ای که اشیاء و اندازه‌گیری‌های فاصله بین آن‌ها را توصیف می‌کند، محاسبه می‌شوند؛ خوشه‌بندی در زمینه‌های مختلفی از جمله امنیت، فیلترهای هرزنامه، هوش تجاری، زیست‌شناسی و جستجوی وب مورد استفاده قرار می‌گیرد [۸].

۵- کارهای مرتبط

◄ تعداد زیادی از محققین علاقه‌مند به تشخیص افسردگی در زمینه‌های مختلف هستند؛ تحقیقاتی در زمینه‌های روان‌شناسی، پزشکی و اجتماعی- زبانی برای شناسایی و ارتباط اختلال افسردگی اساسی و علائم آن انجام شده است؛ علاوه بر این، مطالعات مربوط به الگوریتم‌های داده‌کاوی می‌تواند ویژگی‌های علائم افسردگی را برای بررسی احتمال افسردگی در شبکه‌های رسانه‌های اجتماعی و تجزیه و تحلیل هرگونه نشانه‌ای از افسردگی از طریق پست‌ها، احساسات، گویش و تحلیل احساسات افراد، تجزیه و تحلیل و تشخیص دهد. در این بخش، به تحقیق و روش‌های مربوط به شناسایی و تشخیص افسردگی در شبکه‌های اجتماعی پرداخته می‌شود. اسلام و همکاران [۱۰] از داده‌های فیس‌بوک از یک منبع آنلاین عمومی برای تجزیه و تحلیل افسردگی استفاده کردند؛ در مدل پیشنهادی، از گروهی از ویژگی‌های روان‌زبانی استفاده شد؛ روش‌های یادگیری ماشینی به-عنوان، یک تمرین موثر و مقیاس‌پذیر به‌کار گرفته شد؛ آن‌ها از چهار طبقه‌بندی‌کننده محبوب استفاده کردند: درخت تصمیم، مجموعه، ماشین بردار پشتیبان و k نزدیک‌ترین همسایه؛ از نظر دقت، درخت تصمیم از سایر الگوریتم‌های یادگیری ماشینی که برای ارزیابی نظرات فیس‌بوک برای تشخیص افسردگی استفاده می‌شوند، پیشی می‌گیرد؛ بر اساس فید فعالیت توییتر در طول یک سال، De Choudhury و همکارانش [۴ و ۲۷] از جمع‌سپاری برای جمع‌آوری گروهی از کاربران توییتر که دارای افسردگی بالینی تشخیص داده شده بودند، استفاده کردند؛ آن‌ها صفات رفتاری مربوط به فعالیت اجتماعی، زبان، احساسات و الگوهای زبانی را اندازه‌گیری کردند تا یک طبقه‌بندی آماری ایجاد کنند که می‌تواند خطر افسردگی را قبل از گزارش شروع آن شناسایی کند؛ آن‌ها از طبقه‌بندی‌کننده SVM برای پیش‌بینی قبل از گزارش افسرده شدن یک فرد و احتمال وقوع افسردگی استفاده کردند؛ طبقه‌بندی‌کننده، با دقت طبقه‌بندی ٪۷۰ به نتایجی دست یافت؛ این مطالعه نشان داد که کاربران افسرده کاهش فعالیت اجتماعی، احساسات منفی بیشتر، تمرکز بیشتر بر توجه به خود، افزایش نگرانی های ارتباطی و پزشکی و افزایش بیان عقاید مذهبی را نشان می‌دهند.

دایمی و همکاران [۳]، با استفاده از یک رویکرد مبتنی بر طبقه‌بندی برای پیش‌بینی اینکه کدام بیماران به‌طور بالقوه افسرده هستند، یا قبلا افسرده شده‌اند، استفاده نمودند؛ مدل طبقه‌بندی با استفاده از داده‌های مصنوعی، آموزش و آزمایش شد؛ علائم بر اساس نظرسنجی و مصاحبه با کارشناسان افسردگی انتخاب شدند؛ در این مطالعه، از تکنیک درخت تصمیم C4.5 و ابزار WEKA استفاده شد؛ نتایج نهایی مجموعه داده‌های مصنوعی از نظر دقت، دقت و یادآوری (حساسیت) معقول بود.

رسنیک و همکاران [۱۸ و ۲۷]، با استفاده از سیستم‌های موضوعی نظارت شده به‌منظور تجزیه و تحلیل سیگنال‌های زبانی و برای شناسایی افسردگی، نوعی الگوریتم جدید را مورد بررسی قرار دادند؛ با استفاده از مدل‌های پیشرفته‌تر برای شناسایی و تشخیص افسردگی، مثال‌های کیفی تایید کرده‌اند که مدل LDA، یک تکنیک رایج استخراج موضوع در یادگیری ماشین، می‌تواند ساختار نهفته مهم و بالقوه مفید را با نشان دادن نتایج خوب با استفاده از LDA نظارت‌شده[۲]آشکار کند؛ مدل‌های موضوعی لنگر (SANCHOR)، و همچنین شروع یک کاوش اولیه از یک مدل جدید LDA تودرتو تحت نظارت (SNLDA) برای اولین بار توسط Coppersmith و

[۱] SVMs

[۲] SLDA

همکارانش ارائه شد؛ این دانشمند یک مجموعه داده آزمایشی از توییتر ایجاد کرد که شامل ۳ میلیون توییت از تقریبا ۲۰۰۰ کاربر توییتر بود؛ این‌ها توسط یک روان‌پزشک بالینی واجد شرایط بررسی شد تا مشخص شود که کدام موضوعات بیشتر در ارزیابی افسردگی مرتبط هستند؛ طبق آزمایش کمی، مدل‌های موضوعی پیچیده‌تر با استفاده از نظارت، مانند SLDA و SANCHOR، می‌توانند به تنهایی در LDA بهینه شوند.

برای ارزیابی خطر افسردگی یک فرد، ندیم و همکاران [۱۶] از مجموعه‌ای از ۲,۵ میلیون توییت استفاده کردند تا با در نظر گرفتن رسانه‌های اجتماعی به‌عنوان یک مشکل طبقه‌بندی متن، روش جدیدی برای ساخت یک طبقه‌بندی ایجاد کنند؛ درخت‌های تصمیم، رگرسیون لجستیک، طبقه‌بندی‌کننده بردار پشتیبان خطی و الگوریتم‌های ساده بیز همگی استفاده شدند؛ الگوریتم Naive Bayes با امتیاز ROC AUC ۰,۹۴ درجه گرید A، امتیاز دقت ٪۸۲ و دقت ٪۸۶، از همه طبقه‌بندی‌کننده‌های دیگر پیشی گرفت؛ بنابراین، بهترین مدل برای پیش‌بینی وضعیت سلامت روان کاربر در نظر گرفته شد.

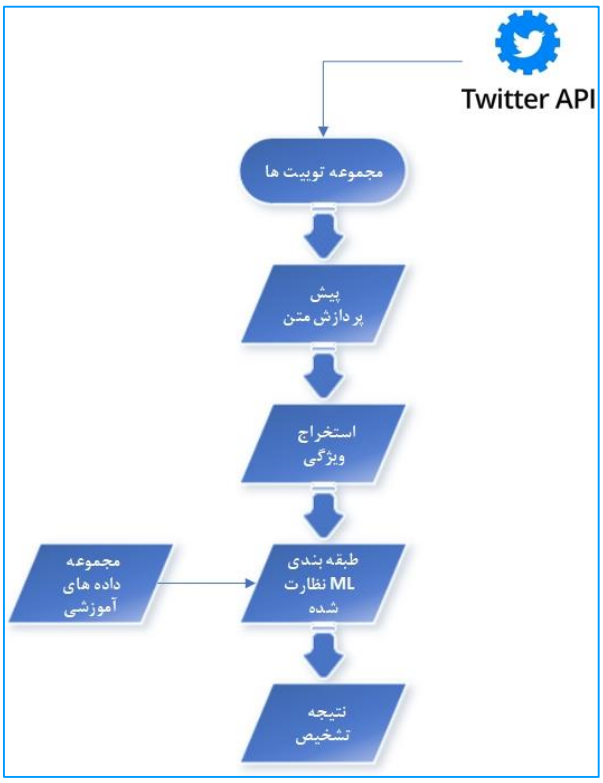
سوناوان و همکاران [۲۱]، یک برنامه کاربردی وب ایجاد کردند که پست‌های رسانه‌های اجتماعی و تست‌های پرسش‌نامه را به‌عنوان ورودی می‌گیرد و سطوح افسردگی را بر اساس خروجی پیش‌بینی می‌کند؛ آن‌ها از طبقه‌بندی‌کننده Naive Bayes (NB) استفاده کردند؛ این سیستم می‌تواند بر اساس پست فیس‌بوک کاربر و همچنین انواع پرسش‌نامه‌های پشتیبانی شده توسط سیستم تشخیص دهد که آیا کاربر استرس دارد یا خیر، تا مکان پزشک مناسب را در نزدیکی مکان کاربر ارائه دهد.

تادسه و همکاران [۲۲ و ۲۷]، پست‌های کاربران را از Reddit مطالعه کردند تا هرگونه ویژگی افسردگی را در کاربران آنلاین کشف کند؛ آن‌ها از یک توصیف فنی از رویکردهای اعمال شده برای تعیین افسردگی با استفاده از روش‌های یادگیری ماشینی و پردازش زبان طبیعی برای طبقه‌بندی متن استفاده کردند؛ آن‌ها یک واژه نامه از اکثر کلمات رایج در حساب های افسردگی ساختند؛ برای استخراج ویژگی‌ها از کاربرد زبانی کاربران در پست‌هایشان، آن‌ها از فرهنگ لغت LIWC، موضوعات LDA و ویژگی‌های N-gram استفاده کردند؛ چارچوب پیشنهادی با استفاده از رگرسیون لجستیک، SVM، جنگل تصادفی، تقویت تطبیقی و طبقه‌بندی‌کننده‌های پرسپترون چندلایه توسعه‌یافته است؛ نتایج نشان داد که طبقه‌بندی‌کننده MLP به دقت ٪۹۱ رسید و بالاترین نتیجه را برای تعیین وجود افسردگی در Reddit با LIWC + LDA + bigram به‌دست آورد؛ علاوه بر این، بهترین ویژگی در میان مجموعه‌های تک ویژگی، بیگرام، طبقه‌بندی کننده SVM با دقت ٪۸۰ است.

آنگ لیا و همکاران [۱۲]، مجموعه داده‌ای از ۱۵۸۷۹ پست را از یک شبکه رسانه اجتماعی چینی به‌نام Weibo تجزیه و تحلیل کردند؛ برای مطالعه از رگرسیون لجستیک ساده، جنگل تصادفی، ماشین‌های بردار پشتیبان و شبکه‌های عصبی پرسپترون چندلایه استفاده شد؛ دو سیستم طبقه‌بندی بر اساس ویژگی‌های زبانی توسعه داده شد: یکی برای تمایز بین پست‌های افسرده و غیرافسرده (انگ/ننگ) و دیگری برای تمایز بین پست‌هایی با سه نوع مختلف انگ افسردگی (غیرقابل پیش‌بینی/ ضعف/ بیماری کاذب)؛ نتایج نشان داد که جنگل تصادفی، بالاترین مقدار F-Measure را برای تمایز بین کلاله / غیر کلاله در ٪۷۵,۲ نشان داد و ارزش تمایز بین سه نوع مختلف کلاله افسردگی ٪۸۶,۲ بود؛ آن‌ها ضرایب شاخص‌ها را در مدل رگرسیون لجستیک ساده (SLR) برای بررسی این روابط تخمین زدند، زیرا ضرایب پیش‌بینی‌کننده‌ها در سه مدل از چهار مدل (مدل‌های SVM، RF و MLPNN) نمی‌توانند به‌وضوح روابط بین ویژگی‌های زبانی و انگ افسردگی را نشان دهند [۲۸].

۶- روش‌شناسی

◄ در این بخش، به روش‌شناسی مدل و مراحل آن پرداخته می‌شود. ما مدلی را مطالعه می‌کنیم که می‌تواند توییت‌های عربی را بر اساس ویژگی‌های افسردگی که توسط متخصصان سلامت انتخاب می‌شود، شناسایی و طبقه‌بندی کند؛ سپس، تکنیک‌های طبقه‌بندی اعمالی بررسی می‌شوند؛ توییت‌ها بر اساس بیشترین ویژگی‌های افسردگی استخراج می‌شوند؛ روش‌های طبقه‌بندی مانند درخت‌های تصمیم، جنگل‌های تصادفی، k نزدیک‌ترین همسایه (KNN)، چندجمله‌ای Naive Bayes (MNB) و ماشین‌های بردار پشتیبان (SVMs) برای این نوع طبقه‌بندی مناسب هستند. همان‌طور که در شکل ۱ نشان داده شده است، مدل شامل مراحل زیر است [۲۷]:



شکل ۱- مروری بر روش تجزیه و تحلیل داده‌های توییتر برای تشخیص افسردگی

۷- جمع‌آوری داده‌ها

◄ مجموعه‌ای از ۳۴۲۴ توییت استخراج شده از شبکه اجتماعی توییتر جمع‌آوری شده است؛ یک اتصال به API توییتر برای جمع‌آوری توییت‌های عربی ایجاد شده است؛ از پست‌های توییتر جمع‌آوری‌شده از API توییتر برای بررسی و تشخیص رفتار افسردگی استفاده شده است؛ مجموعه داده شامل زبان عربی و لهجه‌های سعودی است؛ توییت‌های جمع‌آوری شده باید حاوی نوعی احساسات غم‌انگیز و افسرده باشد و هدف پروژه، استخراج اطلاعات ارزشمند از این توییت‌ها برای شناسایی افسردگی کاربران توییتر است.

۸- پیش‌پردازش متن

◄ چندین مرحله پیش‌پردازش باید روی توییت‌های جمع‌آوری شده انجام شود تا فرآیند تشخیص افسردگی موثرتر شود؛ این مراحل به شرح زیر است [۲ و ۲۷]:

• پاکسازی داده‌ها: شامل مدیریت داده‌های از دست رفته و داده‌های پر سر و صدا می‌شود و نمادهای مختلف مانند علامت تعجب، علائم نگارشی، ارقام و هشتگ‌ها (!، $، ٪، و، #، و غیره) باید حذف شوند؛ علاوه بر این، غلط املایی اصلاح می‌شود و حروف تکراری و توییت‌های غیر عربی حذف می‌شوند [۱۶ و ۱۱].

• حذف کلمه توقف: حذف کلمه توقف عربی، مانند (إلی، من، فی،....) به‌نوعی حذف تمام اشکال توقف کلمات به‌شمار می‌رود؛ برخی از کلمات توقف می‌توانند به دست‌یابی به‌معنای کامل توییت کمک کنند، و برخی از آن‌ها فقط کاراکترهای اضافی هستند که باید حذف شوند.

• عادی‌سازی: عادی‌سازی توییت‌ها مورد نیاز است؛ به‌عنوان مثال، حروف (أ) یا (ا) را می‌توان در حرف اول یک کلمه به‌کار برد؛ برای کلمه‌هایی مانند آحزان، أحزان، آن را به شکل نرمال احزان می‌توان تبدیل‌کرد [۲۷ و ۱۸].

• Tokenization: Tokenization یک فرآیند مهم است؛ تنوع تایپی کلمات را کاهش می‌دهد؛ استخراج ویژگی‌ها نیاز به نشانه‌گذاری دارد؛ پرداختن به زبان عربی به یک جزء سطح بالا نیاز دارد که از یک فرهنگ لغت ویژگی استفاده می‌کند که این کلمات را به بردارهای ویژگی تبدیل می‌کند؛ بنابراین شاخص ویژگی (کلمه) در واژگان با فراوانی آن در کل مجموعه آموزشی مرتبط است.

• ریشه: کلمات موجود در توییت‌ها با حذف هرگونه پسوند، پیشوند و پسوند پیوست شده، ریشه می‌گیرند؛ این کمک می‌کند تا کلمات مشتق شده یا عطف شده را در شکل ریشه، پایه یا ریشه آن‌ها به حداقل برسانید تا فرآیند طبقه‌بندی بهبود یابد؛ به‌عنوان مثال، کلمات حزینه، حزین، احزان همگی ریشه کلمه حزن هستند [۱۹ و ۱۰ و ۲۷].

۹- استخراج ویژگی

◄ پس از پیش‌پردازش متن، مجموعه داده جمع‌آوری شده برای استخراج ویژگی یا ویژگی‌هایی استفاده می‌شود که برای آموزش مدل طبقه‌بندی‌کننده ما استفاده می‌شود؛ انتخاب ویژگی، به افزایش دقت طبقه‌بندی با حذف عبارت‌های نادر از مجموعه‌های آموزشی و آزمایشی توییت‌ها کمک می‌کند؛ هدف استخراج ویژگی، استخراج محتوای مهم یک توییت، استخراج کلمه ویژگی است که پیامی را برای کاربر حمل می‌کند، صرف نظر از اینکه توییتی افسرده‌کننده باشد یا نه [۲۰ و ۲۷]؛ الگوریتم‌های یادگیری ماشینی نیازمند نمایش ویژگی‌ها یا ویژگی‌های کلیدی داده‌ها برای پردازش هستند تا ابعاد و فضای ویژگی‌ها را کاهش دهند و عملکرد طبقه‌بندی‌کننده‌ها را افزایش دهند؛ انتخاب ویژگی‌های عمومی و روش‌های استخراج ممکن است برای تشخیص موارد پرت استفاده شود [۸]؛ با این حال، استخراج ویژگی‌ها همیشه آسان نیست؛ انتخاب ویژگی‌ها یا ویژگی‌های مرتبط با افسردگی مهم‌ترین مرحله برای ساخت مدل طبقه‌بندی است؛ علائم بر اساس نظرسنجی و مصاحبه با متخصصان در زمینه افسردگی برای انتخاب ویژگی‌های مورد نیاز برای طبقه‌بندی

افسردگی انتخاب شدند؛ نمونه‌ای که مجموعه نهایی ویژگی‌ها را نشان می‌دهد، جدول صفحه بعد با ترجمه توییت‌های زبان عربی و گویش سعودی [۳ و ۲۷] ارائه شده است.

۱۰- الگوریتم‌های طبقه‌بندی یادگیری ماشین

◄ در یادگیری ماشینی نظارت شده، کلاس‌های از پیش تعریف‌شده می‌سازیم، توییت‌ها را حاشیه‌نویسی می‌کنیم و برای آموزش طبقه‌بندی‌کننده، آن‌ها را برچسب‌گذاری می‌کنیم؛ پس از تکمیل فرآیند جمع‌آوری داده‌ها، مجموعه داده‌ها را به ٪۸۰ آموزش و ٪۲۰ مجموعه‌های تست تقسیم می‌کنند؛ مجموعه آموزشی، شامل ویژگی‌های ورودی و برچسب‌های کلاس مربوط به آن‌ها است؛ با استفاده از این مجموعه آموزشی، مدل طبقه‌بندی، توسعه می‌یابد که تلاش می‌کند ویژگی‌های ورودی را در برچسب‌های کلاس مربوطه دسته‌بندی کند؛ سپس، مدل با پیش‌بینی برچسب‌های کلاس ویژگی‌های دیده نشده با استفاده از مجموعه آزمایشی، اعتبارسنجی می‌شود؛ در این کار، از الگوریتم‌های یادگیری ماشین نظارت‌شده مانند ماشین‌های بردار پشتیبان (SVM)، k-نزدیک‌ترین همسایه (KNN)، چند جمله‌ای Naive Bayes (MNB)، درخت‌های تصمیم‌گیری و جنگل‌های تصادفی استفاده شده است که برای این نوع طبقه‌بندی مناسب هستند [۲۱ و ۱۴].

جدول ۱- ویژگی‌های افسردگی [۲۷]

لهجه سعودی	ترجمه عربی	ویژگی ها
حزنان, حزینة, حزین	الحزن	غمگینی
الحیاة, خسارة, متعبة, قاسیة	فقدان الاهتمام فی الحیاة	از دست دادن علاقه به زندگی
ملل هالحزه, ما اقدراسوی شی, کسل, خمول	کسل	تنبل
مستقبل مخیف, یوم تعیس, یوم اسود, مافیه امل, مانی متفائل	تشاؤم	بدبینی
انا شین, انا شینة, موحلو, مو حلوه, قبیح, قبیحة	ضعف الصورة الذاتیة	تصویر ضعیف از خود
بموت, ودی اموت	الدافع الانتحاری	انگیزه خودکشی
وش هالاهل, ماعندی اصداقاء, اصحاب منافقین	فقدان الشعور الدافئ تجاه الأسرة أو الأصدقاء	از دست دادن احساس گرما نسبت به خانواده یا دوستان
انسی دائما, انا غبی, غبیة	ذاکرة ضعیفة	حافظه ضعیف
فاشل, فاشلة, ما انجح بشیء, بجیب العید	فشل الماضی	شکست گذشته
ابکی دائما, اصیح, دموع	نوبات البکاء	طلسم گریه
صدمة حیاتی, ضرب, تعنیف الصغار	صدمة الطفولة	ترومای دوران کودکی

چند جمله‌ای ساده و بی‌تکلف (MNB) یک نوع تخصصی از بیزهای ساده (NB) است که برای اسناد متنی با در نظر گرفتن وقوع کلمات در اسناد آموزشی از کلاس طراحی شده است؛ این یک سند متنی نسخه تخصصی ریاضی‌دان ساده‌لوح است؛ به‌دلیل وجود و عدم وجود کلمات صریح، Naive Bayes برای مدل‌سازی یک سند ساده در نظر گرفته می‌شود؛ طبقه‌بندی Simple Naive Bayes از یک سند از وجود و عدم وجود کلمات صریح پشتیبانی می‌کند؛ با این حال، چند جمله‌ای Naive Bayes به صراحت تعداد کلمات را مدل می‌کند [۱۵]؛ علاوه بر این، مدل چند جمله‌ای، اطلاعات فراوانی کلمه را در اسناد انتخاب می‌کند [۱۴ و ۲۷].

ماشین‌های بردار پشتیبان (SVM)، یک مدل پیش‌بینی هستند که از ویژگی‌ها یا متغیرهایی استفاده می‌کنند که از داده‌ها استخراج می‌شوند؛ سپس، ویژگی‌ها به‌عنوان متغیرهای مستقل در یک الگوریتم برای پیش‌بینی متغیر وابسته یک نتیجه مورد علاقه [۶] در نظر گرفته می‌شوند؛ طبقه‌بندی‌کننده SVM یک طبقه‌بندی‌کننده باینری غیر احتمالی است که صفحه جداکننده بین دو کلاس را با حداکثر حاشیه پیدا می‌کند [۱۷ و ۲۸]؛ برخلاف Naive Bayes، طبقه‌بندی‌کننده SVM بردارهای باینری غیر احتمالی را به یک الگوریتم یادگیری برای طبقه‌بندی تبدیل می‌کند؛ این ویژگی‌ها را به‌عنوان نقاطی در فضای پیش‌بینی شده برای یکی از کلاس‌های اختصاص داده شده نشان می‌دهد [۲]؛ طبقه‌بندی‌کننده SVM، از حاشیه بزرگی برای طبقه‌بندی استفاده می‌کند؛ این توئیت‌ها را با استفاده از یک هایپرپلین جدا می‌کند؛ t یک ابر صفحه ایجاد می‌کند که داده‌ها را به دو مجموعه با حداکثر حاشیه در طبقه‌بندی خطی

تقسیم می‌کند؛ هنگامی‌که دو ضلع برابر باشند، یک ابر صفحه با حداکثر حاشیه دارای فاصله‌هایی از ابر صفحه تا نقاط است [۱ و ۲۷]. K نزدیکترین همسایه (KNN) به‌عنوان ساده‌ترین و اساسی‌ترین روش طبقه‌بندی زمانی در نظر گرفته می‌شود که دانش قبلی در مورد توزیع داده وجود نداشته باشد یا به مقدار کمی وجود داشته باشد؛ این قانون، به‌سادگی کل مجموعه داده آموزشی را در حین یادگیری نگه می‌دارد و یک کلاس را بر اساس برچسب اکثر k نزدیک‌ترین همسایگان آن در مجموعه آموزشی به هر کوئری اختصاص می‌دهد؛ اکثریت رای در میان رکوردهای داده در همسایگی معمولا برای تعیین طبقه‌بندی رکورد داده t با یا بدون در نظر گرفتن وزن‌دهی مبتنی بر فاصله استفاده می‌شود [۹]؛ با این حال، برای پیاده‌سازی (KNN)، باید یک مقدار مناسب برای k انتخاب کرد و موفقیت طبقه‌بندی تا حد زیادی، به مقدار k بستگی دارد؛ بنابراین، روش KNN با مقدار k [۷] بایاس می‌شود؛ در مدل، مقدار k، ۳ انتخاب شده است. درخت‌های تصمیم معمولا در تکنیک‌های یادگیری ماشین استفاده می‌شوند؛ آن‌ها به‌سادگی دنباله‌ای از سوالات با دقت طراحی شده را در تلاش برای طبقه‌بندی داده‌ها ایجاد می‌کنند؛ درخت‌های تصمیم، طبقه‌بندی‌کننده‌هایی هستند که برچسب‌های کلاس را برای اقلام داده، پیش‌بینی می‌کنند؛ بسیاری از مشکلات علمی نیازمند برچسب‌گذاری اقلام داده با استفاده از یک کلاس خاص بر اساس ویژگی‌های آیتم داده است؛ درختان تصمیم، با تجزیه و تحلیل یک مجموعه داده آموزشی ساخته می‌شوند که برچسب‌های کلاس برای آن‌ها شناخته شده است؛ سپس از آن‌ها برای طبقه‌بندی نمونه‌های داده‌ای که قبلا دیده نشده‌اند، استفاده می‌شود؛ اگر درخت‌های تصمیم با داده‌های با کیفیت بالا آموزش داده شوند، می‌توانند پیش‌بینی‌های بسیار دقیقی ارائه دهند [۱۱]؛ در تکنیک جنگل تصادفی، چندین درخت تصمیم توسط یک الگوریتم ساخت درخت تصادفی، ساخته می‌شوند؛ با گرفتن رایج‌ترین پیش‌بینی در بین درختان، پیش‌بینی‌های گروه حاصل از درختان تصمیم ترکیب می‌شوند؛ حفظ مجموعه‌ای از فرضیه‌های خوب، به‌جای تعهد به یک درخت، احتمال طبقه‌بندی اشتباه یک مثال جدید را با اختصاص دادن کلاس نادرست توسط بسیاری از درختان کاهش می‌دهد [۱۱].

۱۱- نتایج ارزیابی

◄ در این بخش، آزمایش مدل تشخیص افسردگی را با استفاده از معیارهای ارزیابی متن مانند دقت، یادآوری (حساسیت) و امتیاز F مورد بحث قرار داده می‌شود؛ موارد زیر، متداول‌ترین معیارهای مورد استفاده برای طبقه‌بندی باینری بر اساس مقادیر ماتریس سردرگمی هستند [۱۹]، [۲۰]:

دقت: اثربخشی الگوریتم را با نشان دادن احتمال ارزش واقعی برچسب کلاس تقریبی می‌کند؛ علاوه بر این، اثربخشی کل الگوریتم را با رابطه زیر ارزیابی می کند [۲۷]:

(۱)

$$Accuaracy = \frac{t_p + t_n}{t_p + f_p + f_n + t_n}$$

دقت تعداد نمونه‌های مثبتی است که به‌درستی طبقه‌بندی شده‌اند، تقسیم بر تعداد کل نمونه‌هایی که توسط سیستم به‌عنوان مثبت طبقه‌بندی شده‌اند و ارزش پیش‌بینی یک برچسب را بدون در نظر گرفتن مثبت یا منفی بودن آن، با توجه به کلاسی که در آن قرار دارد، تخمین می‌زند؛ علاوه بر این، قدرت پیش‌بینی الگوریتم را با رابطه زیر ارزیابی می‌کند [۱۶ و ۲۲]:

(۲)

$$\Pr ecision = \frac{t_p}{t_p + f_p}$$

یادآوری (حساسیت): تعداد نمونه‌های مثبتی است که به‌درستی طبقه‌بندی شده‌اند، تقسیم بر تعداد کل نمونه‌های مثبت در داده‌ها؛ حساسیت به احتمال تقریبی است که برچسب مثبت (منفی) درست است؛ علاوه بر این، کارایی الگوریتم را در یک کلاس واحد ارزیابی می‌کند.

(۳)

$$\Re call = \frac{t_p}{t_p + f_n} = Sensitivity$$

F-score یا F-measure: ترکیبی از موارد فوق است و روابط بین برچسب‌های داده مثبت و مواردی که توسط طبقه‌بندی‌کننده ارائه می‌شود را نشان می‌دهد؛ F-measure یک اندازه‌گیری ترکیبی است که از الگوریتم‌های با حساسیت بالا استفاده می‌کند و الگوریتم‌های با ویژگی بالاتر را به چالش می‌کشد، جایی‌که F-measure به‌طور یکنواخت در زمانی که $\beta = 1$ است، متعادل می‌باشد [۱۸ و ۲۷].

(۴)

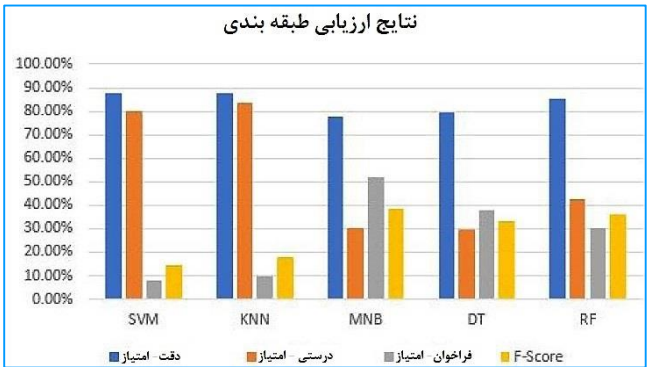
$$F - measure = \frac{(\beta^2 + 1) * \Pr ecision * \Re call}{\beta^2 * \Pr ecision + \Re call}$$

جدول ۲ و شکل ۲، نتایج نهایی ارزیابی را برای الگوریتم‌های اعمال شده برای ماشین بردار پشتیبان (SVM)، k نزدیک‌ترین همسایه

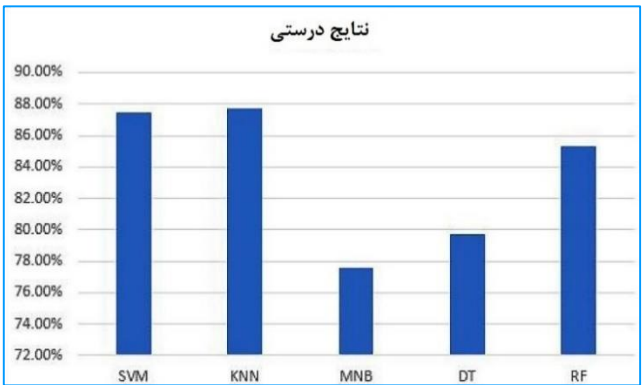
(KNN)، چند جمله‌ای ساده و بی‌تکلف (MNB)، درخت تصمیم (DT) و جنگل تصادفی (RF) برای انتخاب نشان می‌دهد.

جدول ۲- نتایج ارزیابی طبقه‌بندی [۲۷]

ارزیابی	SVM	KNN	MNB	DT	RF
دقت	87.50%	87.70%	77.60%	79.70%	85.30%
درستی	80%	83.30%	30.20%	29.60%	42.80%
فراخوانی	8%	10%	52%	38%	30%
F-Score	14.50%	17.80%	38.20%	33.30%	36.30%



شکل ۲ - نتایج ارزیابی طبقه‌بندی [۲۷ و ۲۸]



شکل ۳- نتایج درستی [۲۷ و ۲۸]

شکل ۳ نشان می‌دهد که طبقه‌بندی‌کننده k نزدیک‌ترین همسایه (KNN) بالاترین دقت را با ٪۸۷,۷۰ دارد؛ به‌دنبال آن ماشین بردار پشتیبان (SVM) با ٪۸۷,۵۰ و جنگل تصادفی (RF) با ٪۸۵,۳۰ قرار دارد. بر اساس نتایج، طبقه‌بندی‌کننده k نزدیک‌ترین همسایه (KNN) تقریبا در تمام معیارها از سایر تکنیک‌های طبقه‌بندی بهتر عمل کرد و پس از آن ماشین بردار پشتیبان (SVM) و جنگل تصادفی (RF) قرار گرفتند [۲۳ و ۲۰ و ۲۷].

۶ - مراجع

۱۲- نتیجه‌گیری

در این مقاله، سیستم پیشنهادی مقالات مبنا مورد بحث قرار گرفت؛ مقالات مربوطه، کارهای قدرتمندی را در حیطه مذکور انجام داده‌اند؛ در مقالات، مدلی طراحی شده است که بتواند توییت‌های عربی را بر اساس ویژگی‌های افسردگی که توسط متخصصان سلامت انتخاب می‌شود، طبقه‌بندی کند؛ پس از آن، تکنیک‌های طبقه‌بندی، به‌کار گرفته شده است که در آن توییت‌ها بر اساس بیشترین ویژگی‌های افسردگی استخراج می‌شوند؛ سپس، دقت تکنیک‌های یادگیری ماشین تحت نظارت اعمال شده، برای انتخاب بهترین الگوریتم برای مدل مربوطه ارزیابی شده است؛ نتایج ارزیابی‌ها نشان داد که طبقه‌بندی‌کننده k-نزدیک‌ترین همسایه (KNN) تقریبا در همه معیارها از سایر تکنیک‌های طبقه‌بندی بهتر عمل می‌کند و پس از آن ماشین بردار پشتیبان (SVM) و جنگل تصادفی (RF) قرار دارند؛ در کار آینده، امیدواریم درک شود که چگونه تحلیل رفتار رسانه‌های اجتماعی می‌تواند به توسعه تکنیک‌های مقیاس‌پذیر گسترده برای ردیابی خودکار سلامت عمومی در جهان عرب کمک کند؛ علاوه بر این، ما علاقمند به استفاده از پتانسیل پلتفرم‌های رسانه‌های اجتماعی برای ردیابی دقیق سلامت روان در جمعیت عربی و شناسایی افراد مبتلا به هدف پیشگیری و ارسال پیام‌های آگاهی توسط پزشکان برای افزایش سلامت روان هستیم؛ همچنین علاقمند به استفاده از تکنیک‌های تکاملی بیشتری در متن عربی برای استخراج ویژگی‌های احساسی بیشتر با بیشترین ویژگی‌های افسردگی هستیم که می‌تواند به بهبود نتایج کمک کند. به‌دلیل پیچیدگی زبان عربی و کمبود منابع و ابزارهای موجود برای استخراج احساسات عربی، به مجموعه داده‌های عربی بیشتری برای تایید اثربخشی و کارایی مدل نیاز است؛ علاوه بر این، گویش‌های عربی در طول زمان در حال تغییر هستند و از ساختار دستوری رسمی عربی استاندارد مدرن، پیروی نمی کنند. مقالات مورد بررسی، روش‌های جذابی را برای مساله مذکور ارائه داده‌اند؛ در آینده تلاش می‌کنیم که ما هم روش‌هایی را به‌کار بگیریم که بتواند کاربردمحور باشد و گامی در جهت توسعه بردارد.

1- Alsaleem, Saleh (2011) Automated Arabic Text Categorization Using SVM and NB. Int. Arab J. e-Technol. 2: 124-8.

2- Atoum, Jalal Omer, Mais Nouman (2019) Sentiment analysis of Arabic jordanian dialect tweets. Int. J. Adv. Comput. Sci. Appl 10: 256-62

3- Daimi, Kevin, Shadi Banitaan (2014) Using data mining to predict possible future depression cases. International Journal of Public Health Science (IJPHS) 3: 231-40.

4- De Choudhury, Munmun, Michael Gamon, Scott Counts, and Eric Horvitz (2013) Predicting depression via social media. In Seventh international AAAI conference on weblogs and social media

5- Elbagir, Shihab, Jing Yang (2018) Sentiment Analysis of Twitter Data Using Machine Learning Techniques and Scikit-learn. In Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence 1-5.

6- Guntuku, Sharath Chandra, David B. Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt (2017) Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences 18: 43-9.

7- Guo Gongde, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer (2003) KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, Berlin, Heidelberg. 986-96.

8- Han, Jiawei, Jian Pei, and Micheline Kamber (2011) Data mining: concepts and techniques. Elsevier.

9- Imandoust, Sadegh Bafandeh, and Mohammad Bolandraftar (2013) Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications 3: 605-610.

10- Islam (2018) Depression detection from social network data using machine learning techniques. Health information science and systems 6: 8.

11- Kingsford, Carl, Steven L Salzberg (2008) What are decision trees?. Nature biotechnology 26: 1011-3.

12- Li Ang, Dongdong Jiao, Tingshao Zhu (2018) Detecting depression stigma on social media: A linguistic analysis. Journal of affective disorders 232: 358-62.

13- Lloyd, Seth, Masoud Mohseni, Patrick Rebentrost (2013) Quantum algorithms for supervised and unsupervised machine learning. arXiv preprint arXiv:1307.0411

14- McCallum, Andrew, and Kamal Nigam (1998) A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization 752: 41-8.

15- Mohana R, S Sumathi (2014) Document classification using Multinomial Naïve Bayesian Classifier. International Journal of Science, Engineering and Technology Research: 1557.

16- Nadeem Moin (2016) Identifying depression on Twitter. arXiv preprint arXiv:1607.07384.

17- Nguyen (2017) Using linguistic and topic analysis to classify sub-groups of online depression communities. Multimedia tools and applications 76: 10653-76.

18- Resnik (2015) Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality:99-107.

19- Sokolova, Marina, Guy Lapalme (2009) A systematic analysis of performance measures for classification tasks. Information processing & management 45: 427-37.

20- Sokolova, Marina, Nathalie Japkowicz, and Stan Szpakowicz (2006) Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence, Springer, Berlin, Heidelberg: 1015-21.

21- Sonawane NP (2018) Predicting Depression Level Using Social Media Posts. IJIRSET 7: 6016-9.

22- Tadesse, Michael M, Hongfei Lin, Bo Xu, Liang Yang (2019) Detection of Depression-Related Posts in Reddit Social Media Forum. IEEE Access 7: 44883-93.

23- Tsugawa (2015) Recognizing depression from twitter activity. In Proceedings of the 33rd annual ACM conference on human factors in computing systems 3187-96.

24- Zadeh LA (1965) Fuzzy sets. Information and Control 8: 338-53.

25- Zhang F, J Ignatius, CP Lim, M Goh (2014) A two-stage dynamic group decision making method for processing ordinal information. Knowledge-Based Systems 70: 189-202.

26- Zhou W, Z Xu (2016) Generalized asymmetric linguistic term set and its application to qualitative decision making involving risk appetites. European J Operational Res 254: 610-21.

27- Amjad A Alaskar, Mourad Ykhlef (2021) Depression Detection from Arabic Tweets Using Machine Learning Techniques. Journal of Computer Science and Software Development March 19, 2021, pp 1-10.

28- Baghdadi NA, Malki A, Magdy Balaha H, AbdulAzeem Y, Badawy M, Elhosseini M. (2022). An optimized deep learning approach for suicide

detection through Arabic tweets. PeerJ Comput. Sci. 8:e1070.

29- Smys S, Raj JS. 2021. Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. Journal of Trends in Computer Science and Smart Technology (TCSST) 3(1):24–39.

30- Kemp S. 2022. Digital 2022: global overview report. Available at https://datareportal.com/reports/digital-2022-global-overview-report (accessed 16 April 2022).

31- Liu D, Feng XL, Ahmed F, Shahid M, Guo J. 2022a. Detecting and measuring depression on social media using a machine learning approach: systematic review. JMIR Mental Health 9(3):e27244.

32- Vanam H, Jeberson RRR. 2021. Analysis of Twitter data through big data based sentiment analysis approaches. Materials Today: Proceedings Epub ahead of print 4 January 2021.

33- Zhang T, Schoene AM, Ji S, Ananiadou S. 2022. Natural language processing applied to mental illness detection: a narrative review. npj Digital Medicine 5(1):1–13.

34- Tong L, Liu Z, Jiang Z, Zhou F, Chen L, Lyu J, Zhang X, Zhang Q, Sadka A, Wang Y, Li L, Zhou H. 2022. Cost-sensitive boosting pruning trees for depression detection on twitter. IEEE Transactions on Affective Computing. Epub ahead of print 25 January 2022.

35- Liu Y, Luo X, Zhang M, Tao Z, Liu F. 2022b. Who are there: discover Twitter users and tweets for target area using mention relationship strength and local tweet ratio. Journal of Network and Computer Applications 199(FEB):103302.

36- Islam M, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A. 2018. Depression detection from social network data using machine learning techniques. Health Information Science and Systems 6(1):1–12.

37- Bhushan B, Sharma N. 2021. Transaction privacy preservations for blockchain technology. In: International Conference on Innovative Computing and Communications. Cham: Springer, 377– 393.