تشخیص حس متن با استفاده از روش‌های یادگیری ماشین

دکتر امیر رجبی
دانشگاه آزاد بندرعباس
A_rajabi@yahoo.com

پری دریاکش
دانشگاه آزاد بندرعباس
Pari.daryakesh@yahoo.com

مونا رجبی‌فرد
دانشگاه آزاد بندرعباس
m.rajabi@tci.ir

کمیل آقابابایی
دانشگاه آزاد بندرعباس
babaiekomeil@gmail.com

مریم یوسفی‌زاده
دانشگاه آزاد بندرعباس
m-yousefizadeh@irimo.ir

چکیده

در متن کاوی به منظور تشخیص احساس متن برای دسته بندی و بررسی نظرها در سایت‌های اجتماعی به منظور تشخیص قطبیت متن، مطالعاتی انجام شده است. هدف این مقاله، بررسی تأثیر الگوریتم‌های طبقه‌بندی برای آنالیز حس متن می‌باشد. در این مقاله ، آنالیز احساس به روش نظارت شده مبتنی بر دادهای برچسب‌گذاری انجام شده است. به‌کارگیری الگوریتم‌های متفاوت در میزان تشخیص حس جمله می‌تواند بسیار موثر باشد. از این رو این مقاله، با در نظر گرفتن الگوریتم‌های طبقه‌بندی جنگل های تصادفی، نایو بیز، درخت اضافی و SVC خطی بررسی و نوشته شده است.

بررسی میزان تاثیر با استفاده از معیارهای Recall و Precision و F-score سنجیده می‌شود. نتایج به دست آمده از اجرای الگوریتم ها بر روی داده های آزمون، سنجیده خواهند شد. بررسی نتایج نشان می‌دهد انتخاب الگوریتمRandom Forest می‌تواند، میزان دقت را افزایش دهد.

واژه‌های کلیدی: یادگیری ماشین , تشخیص احساسات , طبقه‌بند Linear SVC, طبقه‌بند Naive Bayes، طبقه‌بند Random Forest, طبقه‌بند Extra Trees , Bag of Words , TF-IDF

۱. مقدمه

امروزه آنالیز احساس در داده‌های متنی، زمینهٔ مطالعاتی است که سعی در بیان احساس ها، رفتارها، نظرها و تحلیل افراد مختلف نسبت به موجودیت‌ها و ویژگی‌های آن دارد . این موجودیت می تواند محصول، سرویس، سازمان، فرد ، رخداد و موضوع باشد . هدف از آنالیز یا تحلیل احساسات، پیداکردن نظرهایی است که احساسی را نشان داده و جهت گیری این نظرها را تشخیص می دهد. با رشد رسانه های اجتماعی مانند نظرسنجی ها، فروم ها، انجمن های گفت وگو، وبلاگ ها، توییتر و شبکه های اجتماعی، اهمیت آنالیز احساسات افزایش یافته است . سیستم های آنالیز احساسات کمابیش در همهٔ زمینه های تجاری و اجتماعی به کار گرفته می شوند؛ زیرا نظرها و عقیده ها برای همهٔ فعالیت های انسانی مهم بوده و تأثیر شایان توجهی بر رفتار ما دارند.

در این مقاله هدف ، یافتن نگرش متن نظر با توجه به دیدگاه و موضوعات مطرح شده در متن خبر است. در این راستا مجموعه داده ای متشکل از اخبار و نظرات جمع آوری شده و توسط افراد خبره برچسب زده شده و روشی پیشنهادی با استفاده از روشهای متن کاوی ارائه شده است. این روش بر مبنای تحلیل ساختار خبر، یافتن ارتباط بین متن نظر و خبر میباشد. این ارتباط از تجزیه ساختار دستوری متن و استخراج اسامی برای یافتن موضوع متن خبر و سپس یافتن اشتراک در متن نظر و خبر بدست می آید. ساختار این مقاله به شکل زیر است. پس از مقدمه در بخش ۲ نگاهی بر مطالعات انجام شده در سال های اخیر داریم و در بخش ۳ با تشریح مجموعه داده و در بخش ۵ به مقایسه تاثیر الگوریتم های طبقه‌بند می پردازیم.

۲. پیشینه پژوهش

آنالیز احساس در حال حاضر بـه موضـوع روز تبـدیل شده و تحقیقات زیادی در این زمینه انجام شده است . از جمله تحقیقات بر چگونگی استخراج ویژگی و تغییرویژگی ها برای افزایش دقت الگوریتم انجام شده است.

در سال ۲۰۱۸Arxiv JiHoPark با استفاده از مدل های یادگیری ماشین سنتی مانند رگرسیون بردار پشتیبانی (SVR) و رگرسیون منطقی به بررسی تحلیل احساسات متن پرداختند.

در سال ۲۰۱۸ Arxiv ,Christos Baziotisبا استفاده از LSTM دو طرفه با مکانیزم توجه عمیق به پیش بینی محتوای موثر توییت با RNN های عمیق توجه و یادگیری انتقال پرداختند.

در سال ۲۰۱۸ UmangGupta Ankush Chatterjee استفاده از یک مدل یادگیری عمیق مبتنی برLSTM , تشخیص احساسات در مکالمات متنی را بررسی کردند و به کشف احساسات فیزیولوژیکی در حوزه ی علوم رایانه ای پرداختند.

در سال ۲۰۱۸ Ms. Farha Nausheen با جمع آوری توییت کاربران خاص با استفاده از کتابخانه twitter API Twython در پایتون تجزیه و تحلیل احساسات برای پیش بینی نتایج انتخابات پرداخته است.

در سال ۲۰۱۸ Ji Ho Park Arxiv با استفاده از یک مدل CNN بنام EVEC برای مدلسازی احساس درون یک کلمه است که قابلیت ترکیب با کارهای طبقه بندی متن راداراست به پیدا کردن نماینده های خوب از احساسات برای طبقه بندی متن پرداخته است .

در سال ۲۰۱۹ Hakak Nida با استفاده از مدل SVM به طبقه بندی احساسات خودکار پرداخته است.

در سال ۲۰۱۸ Spencer Cappallo با استفاده از شبکه عصبی پیشرفته LSTM و مجموعه ای بزرگ از مقادیر Emojiبرای بیان روابط معنایی بین Emoji و متن/ Emoji و تصاویر به چالش های موجود در پیش بینی ایموجی پرداخته است.

در سال ۲۰۱۸ Toshiki Tomihira با استفاده از مقایسه مدل رمزگذار-رمزگشای شبکه عصبی (RNN)و شبکه عصبی (CNN) به پیش بینی ایموجی عصبی برای تحلیل احساسات پرداخته است.

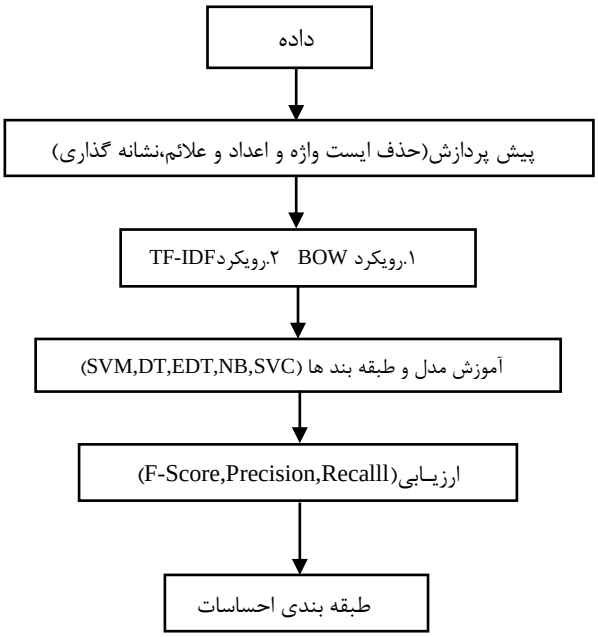
در سال ۲۰۱۹ Bharat Gaind با استفاده از پردازش زبان طبیعی و الگوریتم طبقه بندی Machine Learning به تشخیص احساس و تجزیه و تحلیل در رسانه های اجتماعی پرداخته است.

در سال ۲۰۱۸ Kazuyuki Matsumoto Tokushim با استفاده از دقت بالای روش های مبتنی بر BiLSTM نسبت به روش های مبتنی بر شبکه های عصبی به طبقه بندی دسته های Emoji از Tweet بر اساس شبکه های عصبی عمیق پرداخته است.

۳. روش تحقیق

۳,۱. پیش پردازش

ابزارهای مختلف Stanford CoreNLP [1] برای تجزیه و تحلیل جامع زبانی استفاده می شود. در روش پیشنهاد، ابتدا باید متن ورودی با استفاده از این کتابخانه پیش پردازش شود. در روند اجرای روش پیشنهادی عملیاتی برای آماده سازی متن جهت اجرای روش پیشنهادی نیاز به اعمالی چون حذف ایست واژه ها ،حذف حروف اضافه ،حذف اعداد و نشانه گذاری داریم.



شکل ۱ : روندنمای تحلیل احساسات

۳,۲. داده ها

در آنالیز احساس به غنی بودن داده های انتخاب شده برای طبقه بندی باید توجه کرد [1]. در پژوهش حاضر داده ها از سایت بانک اطلاعاتی اینترنتی فیلم‌ها می باشد و به طور خاص برای آنالیز احساسات انتخاب شده است[۱]. در این سامانه اطلاعات تمام فیلم ها و بررسی آن ها وجود دارد. با آنالیز احساس این بررسی ها ، می توان میزان رضایت‌مندی کاربران از فیلم‌ها و محتوایشان را ارزیابی کرد.

احساسات به صورت باینری مشخص شده‌اند ، به این معنی که بررسی‌هایی که فیلد rating آن ها کمتر از ۵ است، امتیاز احساسی صفر و رکوردهایی که rating آن ها بزرگتر از ۷ است، امتیاز احساسی ۱ را نشان می دهند. در این مجموعه داده هیچ فیلمی بیش از ۳۰ بررسی ندارد. ۲۵۰۰ بررسی برچسب گذاری شده در داده آموزشی مشابه هیچ یک از ۲۵۰۰ بررسی فیلم‌ها در داده آزمایشی نیست. علاوه بر این ۵۰۰۰۰ بررسی بدون برچسب می باشند.

۳,۳. نشانه گذاری(Tokenization): در این مرحله نظرات به کلمات مستقلی که توکن نامیده می‌شود ، شکسته می شود[۲].

۳,۴. حذف ایست‌واژه‌ها: ایست واژه‌ها لغاتی هستند که علی‌رغم تکرار فروان در متن، از لحاظ معنایی دارای اهمیت کمی هستند. در این فاز کلمات کم اهمیت تر و یا ایست‌واژه ها از متون مورد پردازش، حذف می‌گردند. در اغلب کاربردهای متن، حذف این کلمات نتایج پردازش را بهبود می دهد. علاوه بر این از آنجا که بیشتر کاربردهای پردازش متن با حجم عظیمی از داده ها رو به رو هستند، حذف این کلمات سبب کاهش بار محاسبات و افزایش سرعت خواهد شد.

---

[۱] IMDB

برای کاوش کردن مجموعه بزرگی از اسناد ضروریست که اسناد پیش پردازش شوند و اطلاعات در یک ساختار دادهای مناسب برای پردازش های بعدی ذخیره شوند. در این زمینه چندین روش وجود دارند که سعی در بهره گیری از ساختار نحوی و معنایی متن دارند.

در بیشتر روش ها، اسناد به صورت مجموعه ای از کلمات نمایش داده می شوند. بیشتر روش های متن کاوی، الگوریتم های کاوش را روی برچسب های نسبت داده شده به هر سند اعمال می کنند. این برچسب ها ممکن است کلمات کلیدی استخراج شده از سند یا فقط لیستی از کلمات در سند مورد نظر باشند. برای نشان دادن کمترین اهمیت یک کلمه در یک سند معمولا از نمایش بردار استفاده می‌شود، برای هر کلمه یک مقدار اهمیت عددی ذخیره می گردد. روش‌های اصلی و مهم موجود که بر اساس این ایده هستند عبارتند از: مدل فضای بردار، مدل احتمالی و مدل منطقی. چون برخی از روش‌های متن کاوی که بیان می‌شوند از مدل فضای برداری استفاده می‌کنند این روش را مختصرا توضیح می‌دهیم.

۳,۵. بردار کیسه کلمات(BOW)

کیف کلمات (BOW) یک مدل در پردازش زبان های طبیعی است که با هدف دسته بندی مستندات و متون استفاده می‌شود. ایده اصلی آن، به این صورت هست که به هر کدام از کلمات یک عدد Unique نسبت می‌دهیم و Feature بدست آمده بر اساس فرکانس تکرار هر کدام از کلمات به دست خواهد آمد. به طور متعارف، مجموعه‌ای از کلمات با استفاده از فرکانس کلمه یا اهمیت به عنوان ویژگی اغلب برای طبقه‌بندی یا بازیابی سند استفاده می‌شود.یکی از مشکلات این روش، انفجار ابعاد است که با افزایش کلمات، بعد هم افزایش می یابد[3].

۳,۶. تکنیک TF-IDF

۳,۶,۱. جدول تکرار کلمات(TF[۲])

روش اصلی یافتن صفحات مرتبط با یک جستجو، روش سنجش تعداد تکرار (TF) یک کلمه است. هر چه که یک کلمه در یک صفحه بیشتر تکرار شده باشد، آن صفحه ارتباط بیشتری با آن کلمه دارد بنابراین اگر کاربر کلمه ای را جستجو کرد، صفحاتی را نمایش خواهیم داد که آن کلمه در آن‌ها بیشتر تکرار شده باشد.

۳,۶,۲.جدول معکوس تکرار در صفحات(IDF)

به ازای هر لغت، باید فرمولی را استفاده کنیم که هر چه تکرار یک لغت در یک کتاب کمتر باشد، به آن امتیاز بیشتری بدهد (رابطه معکوس) مثلاً می توانیم از فرمول N/DF استفاده کنیم که N تعداد کل کتابها و DF تعداد کتابهای حاوی آن لغت است . با این فرمول هر چه یک لغت کمتر تکرار شده باشد، عدد بزرگتری تولید می شود.

۳,۶,۳.جدول TF-IDF

با داشتن جدول TF و IDF ، می توانیم مرتبط بودن یک لغت با یک صفحه را با ضرب این دو در هم نمایش دهیم :

میزان ارتباط لغت با یک صفحه = TF * IDF

[۲] Term Frequency

یعنی هر چه یک لغت در یک متن بیشتر به کار رفته باشد و در سایر متن‌ها خیلی کم به کار رفته باشد، امتیاز آن صفحه برای آن لغت بیشتر می شود که منطقی هم به نظر می رسد.

به هر کلمه در متن یک وزن اختصاص دهیم. با این کار، می‌توانیم اهمیت یک کلمه را در فرآیند مهندسی ویژگی[۳] بهتر شناسایی کرده و در نهایت ویژگی‌های بهتری را برای تزریق به الگوریتم‌های بعدی مانند طبقه‌بندی یا خوشه‌بندی داشته باشیم. مقدار TF-IDF مخفف دو کلمه است: TF به معنیِ Term Frequency یعنی تعداد تکرار یک کلمه در یک متن و عبارت IDF به معنی Inverse Document Frequency که می‌توان آن را به برعکس تعداد تکرار در متون ترجمه کرد.وزن دهی TF-IDF طبق فرمول (۱) محاسبه می شود.

$$f(w) = TF(w).IDF(w) = TF(w).log\frac{N}{n(w)+1} \qquad (۱)$$

۴. الگوریتم های دسته بندی

۴,۱. الگوریتم جنگل تصادفی[۴]

الگوریتم جنگل تصادفی یکی از روش‌های زیر مجموعه درخت های تصمیم گیری است که برای حل مسأله، تعداد زیادی درخت تصادفی روی زیر مجموعه هایی از مجموعه داده تولید می نماید[4]و دقت طبقه بندی این روش با ساخت مجموعه ای از درختان و رأی گیری بین آن ها برای به دست آوردن رده ای با بیشترین تعداد رأی، پیشرفت های قابل توجهی داشته است[5]. جنگل تصادفی یک روش یادگیری ماشین چند منظوره است که قادر به انجام هر دو وظیفه رگرسیون و طبقه‌بندی است. یک نوع از روش یادگیری گروهی است ، که در آن گروهی از مدل‌های ضعیف ترکیب می‌شوند تا یک مدل قدرتمند را شکل دهند[6]. یکی از پارامترهای مهم و موثر در دقت دسته بندی، تعداد درختان در جنگل است. هر یک از درختان موجود در جنگل به تنهایی کارایی زیادی نداشته و در واقع قدرت تشخیص جنگل به برآیند قدرت کلیه ی درختان وابسته است[7] . یکی از مزایای جنگل تصادفی , استفاده از داده‌های بزرگ با ابعاد بالا است . این متد دارای روش‌هایی برای تعادل خطاها در مجموعه داده‌هایی است که در آن کلاس‌ها نامتعادل هستند. جنگل تصادفی شامل نمونه‌برداری از داده ورودی با جایگزینی به نام نمونه‌برداری بوت استرپ[۵] است.

۴,۲. الگوریتم درختان اضافی[۶]

الگوریتم Extra-Trees یک مجموعه ای از درخت تصمیم گیری یا رگرسیون غیرمعمول را ایجاد می کند با توجه به روش کلاسیک بالا به پایین. دو تفاوت اصلی آن ،با سایر روش های گروهی، بر پایه دسته بندی این است که، گره ها را با انتخاب نقطه های برش به صورت تصادفی جدا می کند و از کل نمونه یادگیری (به جای بوت استرپ) برای رشد درخت استفاده می کند[8].

---

[۳] Feature Engineering

[۴] Random Forest Algorithm

[۵] Bootstrap Sampling

[۶] Extremely randomized trees

۴،۳. الگوریتم ماشین بردار پشتیبان[۷]

ماشین بردار پشتیبان (SVM) یک الگوریتم آموزشی است. طبقه‌بند را برای پیش‌بینی نمونه جدید ازکلاس، آموزش می‌دهد. SVM عمدتاً مبتنی بر ایده صفحات تصمیم‌گیری است که به صراحت در مورد مرز تصمیم‌گیری صحبت می‌کند و حوزه‌ای که مرز تصمیم‌گیری بین کلاس‌ها می باشد را به عنوان پارامتر مورد استفاده قرار می‌دهد.

دو کاربرد کلیدی از تکنیک SVM وجود دارد که برنامه‌نویسی ریاضی و تابع کرنل هستند. SVMمی تواند سطح جداکننده را بین داده های کلاس های متفاوت درفضای چند بعدی ، بهینه کند.طبقه بند بردار پشتیبان[۸] (SVC) طبقه بندی هیبرید را جستجو می کند. SVC به عنوان تابع کرنل روی سطوح تصمیم‌گیری غیر خطی کاربرد دارد[9] .

در واقع SVM یک Hyperplane که بین نمونه های مثبت و منفی مجموعه آموزش قرار می گیرد را مشخص می کند. پارامترهای b_j به گونه‌ای تنظیم می شوند که فاصله بین Hyperplane و نزدیک‌ترین نمونه مثبت و منفی ماکزیمم شود.که طبق فرمول (۲) محاسبه می گردد.

$$y = b_0 + \sum_{j=1}^{n} b_j t_{aj} \qquad (۲)$$

۴،۴. الگوریتم نایو بیز[۹]

Naïve Bayes یک الگوریتم یادگیری ماشین برای حل مشکلات طبقه‌بندی است. برای ساخت مدل و پیش بینی سریع از قضیه احتمال استفاده می کند. اساس قضیه Bayes، برپایه استقلال ویژگی‌ها است. یک روش مبنا برای طبقه بندی متن و حل مشکل قضاوت اسناد درباره تعلق یک سند به یک دسته (مانند هرزنامه یا مشروع، ورزش و یا سیاست، و غیره)است که از روش تعیین فرکانس کلمه به عنوان ویژگی استفاده می‌کند. پیش پردازش مناسب، در این الگوریتم با روش‌های پیشرفته تر از جمله ماشین های بردار پشتیبان [10]در رقابت است.

۵. نتیجه و ارزیابی

در این مقاله هدف، بررسی و مقایسه اثر بخشی الگوریتم های داده کاوی ،جهت آنالیز احساس به روش نظارت شده مبتنی بر داده های برچسب‌گذاری انجام شده است . در این راستا از الگوریتم‌های جنگل‌های تصادفی، درخت اضافی، بردار پشتیبان و نایو بیز استفاده شده است که با بررسی چگونگی عملکرد هر الگوریتم و پارامترهای تاثیرگزار آن مورد ارزیابی و بررسی قرار گرفت. در ابتدای این پژوهش، متون با دو رویکرد به فضای برداری انتقال داده می شود.که در رویکرد اول از تکنیک BOW و در رویکرد دوم از TF-IDF استفاده گردید. طبق جدول (۱)، در همه الگوریتم ها، با تکنیک TF-IDF نتایج بهتری حاصل گردید. در ضمن جنگل های تصادفی عملکرد بهتری نشان دادند . دلیل اصلی آن بکارگیری تکنیک ترکیب در این الگوریتم می باشد.

که با استفاده از تعدادی از الگوریتم های درخت تصمیم و تهیه دیتاست های مختلف با تکنیک جایگشتی و در نهایت رأی‌گیری اکثریت ما بین این درختان، نتیجه نهایی بدست آمد. با توجه به کاربرد و اهمیت نتایج حاصل از تحلیل احساسات در حوزه های

---

[۷] Support Vector Machine

[۸] Support Vector Classifier

[۹] Naïve Bayes

مختلف، به‌کارگیری مدل پیشنهادی جنگل‌های تصادفی توصیه شود. درضمن نظرات کاربران، منعکس کننده عقاید و نظرات واقعی آن‌ها در مورد محصولات و خدمات می‌باشد و از این جهت منبع ارزشمندی برای ارائه پیشنهاد به مخاطبین است.

جدول ۱: مقایسه تاثیر الگوریتم ها

Bow				
	جنگل تصادفی	ExtraTrees	Linear SVC	نایو بیز
دقت(p)	۰,۷۸۴۶۹	۰,۷۶۷۶۸	۰,۷۷۳۷۱	۰,۷۷۰۸۶
صحتR	۰,۷۸۰۸۵	۰,۷۶۲۵۴	۰,۷۷۳۶۷	۰,۷۶۱۷۱
معیار F	۰,۷۷۹۷۴	۰,۷۶۰۹۷	۰,۷۷۳۵۹	۰,۷۵۹۱۲
TF-IDF				
	جنگل تصادفی	ExtraTrees	Linear SVC	نایو بیز
دقت(p)	۰,۸۰۵۴۵	۰,۷۷۸۵۰	۰,۷۹۱۵۹	۰,۷۸۹۸۴
صحتR	۰,۸۰۳۳۴	۰,۷۷۲۱۸	۰,۷۹۱۶۰	۰,۷۸۹۸۴
معیار F	۰,۸۰۲۶۶	۰,۷۷۰۴۲	۰,۷۹۱۵۹	۰,۷۸۹۷۹

۶. منابع

[1] C. D. Manning, J. Bauer, J. Finkel, and S. J. Bethard, “The Stanford CoreNLP Natural Language Processing Toolkit,” pp. 55–60, 2014.

[2] T. Tran, D. Nguyen, A. Nguyen, and E. Golen, “Sentiment Analysis of Emoji-based Reactions on Marijuana-Related Topical Posts on Facebook,” *2018 IEEE Int. Conf. Commun.*, pp. 1–6, 2018.

[3] K. Matsumoto, “Classification of Emoji Categories from Tweet Based on Deep Neural Networks,” pp. 17–25, 2018.

[4] E. Scornet, “A random forest guided tour,” vol. 25, no. 2, pp. 197–227, 2016.

[5] ز .ا .خویی ,"طبقه بندی ترافیک شبکه با استفاده از الگوریتم جنگل تصادفی بهبودیافته" ., p. 15, 30 11 1395.

[6] S. RAY, "Powerful Guide to learn Random Forest(With codes in R & Python)," *Powerful Guide to learn Random Forest(With codes in R & Python)*, p. 7, 7 SEPTEMBER 2015.

[7] .م ,مسلمی ," مشخصات نویسندگان مقاله بررسی تعداد درختان در الگوریتم جنگل تصادفی برای تحلیل عقاید در فروشگاههای مجازی"., p. 6, 01 01 1395.

[8] A. Tyryshkina, N. Coraor, and A. Nekrutenko, “Predicting runtimes of bioinformatics tools based on historical data : Five years of Galaxy usage,” pp. 1–10, 2019.

[9] “Immune Support Vector Machine Approach for Credit Card(svm).pdf.” .

[10] A. Varsani and K. Manoj, “Advance and Innovative Research,” no. April, 201