مجتمع آموزش عالی فنی و مهندسی اسفراین
دانشکده مهندسی برق ، کامپیوتر ، صنایع

متن کاوی در متون ادبیات فارسی

پایان‌نامه کارشناسی رشته مهندسی کامپیوتر (گرایش نرم افزار)

نگارنده

علی زاهدی سهروفروزانی

استاد راهنما

دکتر حجت اله موید

بهمن ۱۴۰۲

به نام خداوند جان و خرد

کزین برتر اندیشه برنگذرد

مجتمع آموزش عالی فنی و مهندسی اسفراین
دانشکده مهندسی برق ، کامپیوتر ، صنایع

پایان‌نامه‌ی کارشناسی رشته‌ی مهندسی کامپیوتر آقای علی زاهدی سهروفروزانی

تحت عنوان:

متن کاوی در متون ادبیات فارسی

در تاریخ ..... توسط کمیته‌ی تخصصی زیر مورد بررسی و تایید نهایی قرار گرفت.

۱- استاد راهنما: دکتر حجت اله موید امضاء.....

۲- مدیریت گروه مهندسی کامپیوتر:دکتر سمیه گرایلو امضاء.....

بهمن ۱۴۰۲

قدردانی

با تشکر از استاد راهنمای گرامی، جناب دکتر حجت اله موید، بابت توصیه‌های علمی و تکنیکی خود در طول انجام پروژه به من کمک فراوان کردند و بدون حمایت ایشان، این پروژه به سختی به اتمام می‌رسید.

همچنین، من از دانشکده مهندسی برق کامپیوتر و صنایع، مجتمع آموزش عالی فنی و مهندسی اسفراین قدردانی می‌کنم که با فراهم کردن امکانات لازم و حمایت های علمی، به من در انجام این پروژه کمک کردند.

با سپاس فراوان

کلیه‌ی حقوق مادی و معنوی مربوط به نتایج مطالعات، ابتکارات و نوآوری‌های ناشی از تحقیق موضوع این پایان‌نامه متعلق به "مجتمع آموزش عالی فنی و مهندسی اسفراین" است

تقدیم می کنم به خانواده عزیزم که همواره از حمایت من دست نکشیدند

فهرست مطالب

فهرست مطالب ..... ث

فهرست جداول ..... خ

فهرست اشکال ..... د

چکیده ..... ذ

فصل ۱ مقدمه ۱

۱-۱ پیشینه تحقیق و اهمیت موضوع ۱

۱-۲ هدف از انجام پایان نامه ۲

۱-۳ روش تحقیق ۳

۱-۴ ساختار پایان‌نامه ۴

فصل ۲ مروری بر تحقیقات ۵

۲-۱ مقدمه ۵

۲-۲ چالش‌ها ۶

۲-۳ اهمیت موضوع ۷

۲-۴ روش‌شناسی ۷

۲-۵ جوانب پروژه ۸

فصل ۳ مفاهیم پایه ۱۲

۳-۱ متن‌کاوی ۱۲

۳-۲ پیش‌پردازش ۱۳

۳-۳ کلمات توقف ۱۴

۳-۴ پاک‌سازی دیتاست ۱۵

۳-۵ بازیابی اطلاعات ۱۶

۳-۶ مدل‌های بازیابی اطلاعات ۱۷

۳-۷ مفاهیم پایه بازیابی اطلاعات ۱۸

۳-۸ روش‌ها و مراحل بازیابی اطلاعات ۱۹

۳-۹ کاربردها و دامنه‌های استفاده ۱۹

۳-۱۰ برخی از تکنیک‌های بازیابی اطلاعات ۲۰
۳-۱۱ معیارهای امتیازدهی ۲۲
۳-۱۱-۱ تی اف ۲۲
۳-۱۱-۱-۱ رابطه ریاضی ۲۳
۳-۱۱-۲ دی اف ۲۳
۳-۱۱-۲-۱ رابطه ریاضی ۲۴
۳-۱۱-۳ مقایسه DF و TF ۲۴
۳-۱۱-۴ ای دی اف ۲۵
۳-۱۱-۴-۱ رابطه ریاضی ۲۶
۳-۱۱-۴-۲ مقایسه با معیارهای قبلی ۲۶
۳-۱۱-۵ تی اف ای دی اف ۲۶
۳-۱۱-۵-۱ رابطه ریاضی ۲۸
۳-۱۱-۵-۲ مقایسه با معیارهای قبلی ۲۹
فصل ٤ مطالعات و تحقیقات ۳۰
۴-۱ روند کلی ۳۰
۴-۲ پیش‌پردازش ۳۱
۴-۳ رویکرد ما در عملیات بایس کردن استاپ وردها ۳۳
۴-۴ محاسبه TF-IDF ۳۳
۴-۵ اعلان متغیر ۳۴
۴-۶ مدل ذخیره‌سازی TF-IDF ۳۴
۴-۷ تست داده های جدید ۳۶
۴-۸ روش‌های تست ورودی ۳۶
۴-۹ نحوه تشخیص احساس از TF-IDF ۳۷
۴-۱۰ دقت مدل ۳۷
فصل ٥ نتیجه گیری و پیشنهادات ۳۹

۵-۱ نتیجه گیری ۳۹
۵-۲ پیشنهادات ۴۳
مراجع ٤٦

فهرست جداول

جدول ۴–۱ داده‌های اصلی ۳۱
جدول ۴–۲ فایل‌های پیش پردازش ۳۲
جدول ۴–۳ کلمه بر حسب TF-IDF برای هر شش احساس ۳۵
جدول ۴–۴ نمونه از داده های تست ۳۷
جدول ۴–۵ کد نمونه‌ها ۳۸

فهرست اشکال

شکل ۲–۱ فراوانی سند احساسی..... ۸

چکیده

با توجه به عنوان مسئله و ارتباط مستقیم موضوع با بحث بازیابی اطلاعات صورت مسئله یافتن الگو های زبانی (ترکیبی از کلمات) در یک سند متنی به زبان فارسی است بدین شکل که با دریافت چندین سند متنی از ورودی کاربر پیرامون یک موضوع خاص به دنبال الگو ها می گردد ، این تشخیص الگو به وسیله ابزار های آماری در بازیابی اطلاعات انجام می شود نظیر فراوانی کلمه در سند و مجموعه و دیگر ابزار که در ادامه به آنها اشاره خواهد شد.

پس از تشخیص الگو با در نظر گرفتن روش بهینه، سند دیگری دریافت و اقدام به واکاوی متون از آن می کند که انتظار می رود همان الگویی باشد که در اسناد اولیه پردازش شده یافته بودیم.

اهمیت موضوع از آنجا نمایان می شود که تا کنون پژوهش مشابهی پیرامون آنچه توضیح داده شد در زبان فارسی صورت نگرفته است. از همین رو اهمیت پژوهش دو چندان خواهد بود و منابعی که می توان به آن مراجعه کرد همگی جهت استفاده در زبان انگلیسی و دیگر زبان های روز دنیا خواهند بود که با به پایان رسیدن این پروژه متن کاوی در متون ادبیات فارسی به شکل کاملا عملی و علمی صورت پذیر خواهد بود.

رویکرد ما مبتنی بر قرار دادن فرایند ها و ابزار ها جهت رسیدن به نتایج، جمع آوری و تحلیل و تفسیر می باشد. در این مسیر از روش های کمی و کیفی ، ترکیبی ، تجربی ، توصیفی ، تحلیل استفاده می شود.

جامعه هدف ما ترکیبی از ادبیات معاصر با در نظر داشتن حالات و احساسات و رفتار ها می باشد، نحوه انتخاب نمونه بصورت تصادفی و هدفمند از مجموعه داده ای تحت عنوان "مجموعه ۳۰ هزار جمله ای" که هر جمله از آن را با در نظر گرفتن ۶ احساس اصلی امتیاز بندی شده است استخراج شده از پلتفرم توییتر می باشد.

واژگان کلیدی: بازیابی اطلاعات، فراوانی کلمات، متن کاوی

فصل ۱
مقدمه

۱-۱ پیشینه تحقیق و اهمیت موضوع

متن‌کاوی در ادبیات فارسی به عنوان یک زمینه تحقیقی مهم در حوزه مطالعات ادبیات و پردازش زبان طی دهه‌های اخیر به شدت توجه پژوهشگران جلب کرده است. ادبیات فارسی با بی‌شماری آثار و متونی از دوران کلاسیک تا ادب معاصر، از غنای فراوانی از احساسات و نگرانی‌ها نمایانگر است. با این حال، تشخیص و تحلیل احساسات موجود در این متون برای محققان همچون ادب‌شناسان، متن‌شناسان، و همچنین برنامه‌نویسان کار سختی است.

تحلیل احساسات و نوع جمله در متون ادبیات فارسی اهمیت زیادی در درک عمیق‌تر این متون دارد. ادبیات به‌عنوان یک نمایه از فرهنگ و تاریخ یک جامعه، شامل انواع مختلفی از احساسات و جریانات انسانی است. توانایی تشخیص کمیته‌های احساسی و نوع جمله در متن‌ها، به ادب‌شناسان امکان می‌دهد تا درک بهتری از نوع و مضمون اثرات ادبی داشته باشند.

برای مطالعهٔ تطبیقی ادبیات، تحلیل احساسات و نوع جمله در متون می‌تواند به مقایسهٔ متن‌ها از منظر احساساتی و ادبی کمک کند. این تحلیل می‌تواند به پژوهشگران کمک کند تا تأثیر تاریخ، فرهنگ و محیط اجتماعی بر متون ادبی را بهتر درک کنند.

در کاربردهای کنونی متن‌کاوی و پردازش زبان طیف گسترده‌ای از امکانات فراهم کرده است. از مدل‌های یادگیری عمیق برای تشخیص احساسات در متن تا تحلیل خودکار محتوای ادبی، این پروژه می‌تواند از مزایای پیشرفت‌های فناوری در زمینه پردازش زبان برای تحلیل متون ادبی استفاده کند.

با تأکید بر اهمیت تحلیل احساسات و نوع جمله در متون ادبیات فارسی و توانایی‌های پیشرفته پردازش زبان، این پروژه می‌تواند به افزایش فهم عمیق‌تر از متون ادبی و کاربردهای علمی و فرهنگی آنها در جامعه کمک منجر شود.

۱-۲ هدف از انجام پایان نامه

هدف اصلی این پایان‌نامه ارتقای فهم عمیق‌تر و دقیق‌تر متون ادبیات فارسی با بهره‌گیری از تحلیل احساسات و نوع جمله است. این پروژه باتوجه‌به غنای فراوان از متون ادبی و انواع مختلفی از احساسات و جریانات انسانی در این متون، تلاش می‌کند تا توانایی تشخیص و تحلیل احساسات مختلف موجود در این متون را بهبود بخشد.

از اهمیت‌ها و دلایل می‌توان به موارد زیر اشاره کرد: فهم عمیق‌تر: تحلیل احساسات متون ادبیات فارسی به ما امکان می‌دهد تا فهم عمیق‌تری از طبیعت و مضمون این متون پیدا کنیم. متون ادبیاتی به‌تنهایی تنها کلمات نیستند؛ آنها حاوی احساسات، افکار و تجربیات انسانی عمیقی هستند که با تحلیل دقیق‌تر احساسات متن، می‌توانیم به فهم عمیق‌تری از آثار ادبی دست یابیم.

تحلیل متون مقایسه‌ای: با تحلیل احساسات و نوع جمله در متون ادبیات فارسی، می‌توانیم به تحلیل متون مقایسه‌ای بپردازیم و اثرات تاریخ، فرهنگ و محیط اجتماعی را در متون مقایسه کنیم. این می‌تواند به ما کمک کند تا به مفاهیم عمیق‌تری از تطبیق متون ادبی دست یابیم.

کاربردهای فناوری: در دنیای امروزی، فناوری‌های پردازش زبان و متن‌کاوی پیشرفت‌های چشمگیری داشته‌اند. این پروژه از این تکنولوژی‌ها بهره‌مند شده و به تحلیل خودکار متون ادبی به‌راحتی و دقت کمک می‌کند.

این پایان‌نامه امیدوار است که به پیشرفت‌های درک متون ادبی فارسی و کاربردهای علمی و فرهنگی آنها کمک کند و در راستای بهترین شناخت و تجزیه‌وتحلیل متون ادبی به کمک محققان و علاقه‌مندان در این حوزه باشد.

۱-۳ روش تحقیق

تشخیص احساسات از متون یک مسئله مهم در حوزه پردازش زبان طبیعی و متن‌کاوی است. باتوجه‌به افزایش روزافزون متون موجود در شبکه‌های اجتماعی و پلتفرم‌های آنلاین، توانایی تشخیص احساسات موجود در متون، به ما امکان می‌دهد تا علاوه بر درک مفاهیم متن، نگاهی به جریانات احساسی مرتبط با آن داشته باشیم.

ما از یک مجموعه‌داده ای استفاده می‌کنیم که شامل متون از شبکه‌های اجتماعی به زبان فارسی با برچسب‌های مربوط به شش احساس مختلف (خشم، ترس، خوشحالی، نفرت، غم و تعجب) است. ما از این داده‌ها بهره‌مند می‌شویم تا با استفاده از روش TF-IDF[1] وزن دار وزن‌دهی به کلمات مربوط به هر یک از این احساسات را انجام دهیم[1].

روش TF-IDF یک روش معمول برای بررسی وزن کلمات در یک متن است. با اعمال این روش به داده‌های ما، ما می‌توانیم وزن‌دهی به کلمات مربوط به هر احساس از آن‌ها انجام دهیم. این کار امکان تفکیک کلماتی که بیشترین ارتباط با هر احساس دارند را فراهم می‌کند.

[1] Term Frequency-Inverse Document Frequency

هدف اصلی این است که بتوانیم با استفاده از این معیار، کلمات مربوط به هر احساس را به‌دقت تشخیص داده و از آنها برای تحلیل احساسات در متون مختلف استفاده کنیم. این اطلاعات می‌تواند به ما کمک کند تا درک بهتری از تغییرات احساساتی در متون داشته باشیم و نهایتاً به تفهیم بهتری از مضمون و مفاهیم متون ادبی دست یابیم.

۱-۴ ساختار پایان‌نامه

ابتدا با معرفی موضوع و اهمیت تشخیص احساسات از متون به‌عنوان مسئله اصلی، به معرفی داده‌های مورداستفاده پرداختیم. سپس به‌مرور ادبیات در زمینه تشخیص احساسات و تفسیر مفاهیم مشابه پرداختیم. در مرحله بعد، متدولوژی تحقیق شرح داده شد و مراحل پیش‌پردازش متن، تعیین وزن به کلمات با استفاده از TF-IDF وزن دار و اجرای آزمایش‌ها تشریح گردید. پس از اجرای آزمایش‌ها، نتایج به‌دست‌آمده مورد تجزیه‌وتحلیل قرار گرفتند و با تحلیل مفصل به نتیجه‌گیری رسیدیم. در نهایت، توصیه‌ها و پیشنهادات برای تحقیقات آینده در این زمینه ارائه گردید.

.

فصل ۲
مروری بر تحقیقات

۲-۱ مقدمه

مروری بر اکثر کارهای گذشته نشان می‌دهد که روش‌های مورداستفاده برای سیستم‌های تحلیل احساسات را می‌توان به گروه‌های زیر تقسیم کرد:

- رویکردهای کلیدواژه
- رویکردهای یادگیری ماشینی
- رویکردهای ترکیبی

اولین رویکرد محققان روی شناسایی و استفاده از کلمات کلیدی که مستقیماً احساسات بیان شده در متن را منعکس می‌کنند، کارکرده‌اند. برای شناسایی این کلمات کلیدی از چندین روش استفاده شده است. دقت این روش‌ها به وجود کلمات کلیدی بستگی دارد. یکی از معایب این روش وابستگی آن به مناطق خاص است[2].

عموماً در این رویکرد فهرستی از کلمات با برچسب‌های معنایی عواطف مثبت و منفی یا احساس غم، شادی و... تهیه می‌شود.

گاهی اوقات فهرست با دادن نکاتی ارائه می‌شود که نشان‌دهنده شدت یک احساس است. سپس متن به کلمات جداگانه تقسیم می‌شود. درمرحله بعد، کلماتی که احساسات را پوشش می‌دهند مشخص می‌شود. سپس شدت احساسات اندازه‌گیری می‌شود.

درمرحله بعد بحث می‌شود که آیا فعل منفی است یا خیر. فعل منفی بودن احساس تعیین شده را تغیر می‌دهد. برخلاف رویکردهای مبتنی بر کلیدواژه، رویکردهای مبتنی بر یادگیری ماشینی، با استفاده از سیستم‌های طبقه‌بندی از پیش آموزش‌دیده، سعی درشناسایی احساسات بیان شده در متن دارند. عدم وابستگی به حوزه‌های خاص منجر به پیشرفت سریع روش‌های یادگیری ماشینی تحت نظارت و بدون نظارت شده است.

۲-۲ چالش‌ها

به دلیل کمبود منابع احساسی برچسب‌گذاری شده در دسترس، صفحات توییتر و تگ‌ها مورداستفاده کاربران به‌عنوان منابع پردازشی موجود در بسیاری از مطالعات مورداستفاده قرار گرفته‌اند.

این مجموعه‌داده‌ها می‌توانند به‌عنوان مجموعه‌داده‌های فعال برای پردازش استفاده شوند. آنچه در بیشتر تحقیقات رایج است استفاده از مجموعه‌داده‌ها، ویژگی‌ها و روش‌های طبقه‌بندی مختلف به‌منظور دستیابی به‌دقت بهتر است.

به عنوان مثال ، دریک کار مشابه، دقت ۷۳ درصد با استفاده از الگوریتم ماشین بردار پشتیبان به دست آمده است که از ویژگی هایی مانند کلیدواژه ها معنایی برای روابط معنایی کلیدواژه ها استفاده شده است.

۲-۳ اهمیت موضوع

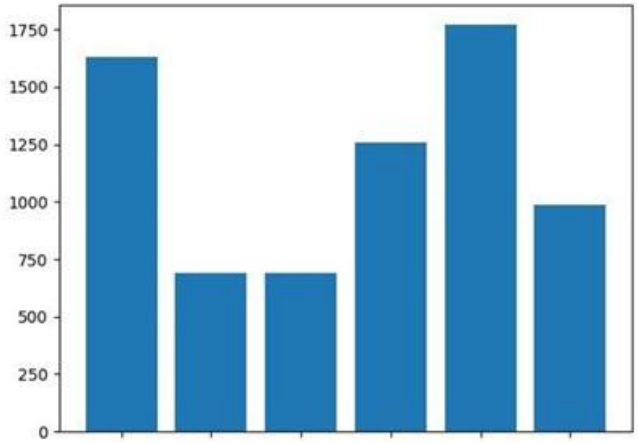
تشخیص احساسات، متن را به یک حوزه مهم برای مطالعه تبدیل می‌کند. باتوجه‌به مشکلات موجود، این کار بسیار چالش‌برانگیز به نظر می‌رسد. به‌عنوان‌مثال، تشخیص احساسات واقعی انسان گاهی اوقات دشوار است به علت اینکه گاهی اوقات افراد از کلماتی استفاده می‌کنند که احساس واقعی‌شان را نشان نمی‌دهد.

نکته دیگر اینکه سبک نوشتن برای هر فرد یکسان نیست و تشخیص مستقیم احساسات واقعی آن‌ها در سبک‌های رسمی و غیررسمی دشوار است. این موارد نمونه‌هایی هستند که تشخیص احساسات را آن طور که نگاه اول به نظر می‌رسد آسان نمی‌کنند.

۲-۴ روش‌شناسی

تعیین نوع احساس بیان شده در متن روش پیشنهادی است. با استفاده از مجموعه توییت‌های رتبه‌بندی شده بر اساس عاطفه، ترس، اندوه، شادی، خشم و حیرت است. در ابتدا، یکی از مشکلاتی که به وجود می‌آید تحلیل احساسات متون تولید شده توسط کاربران در رسانه‌هایی مانند توییتر است. یکی از وظایف اصلی تحلیل احساسات طبقه‌بندی کلمات در دسته‌بندی احساسات می‌باشد.

ارزیابی جملات در منظرهای مختلف برای دریافت احساسات متعددی که یک متن واحد به وجود می‌آیند، تازگی این پروژه است. در تصویر زیر فراوانی جمع کل امتیاز احساسات را به تفکیک مشاهده می‌کنید.



شکل ۲-۱ فراوانی سند احساسی

۲-۵ جوانب پروژه

استفاده از مجموعه‌داده‌های مناسب یکی از عواملی است که بر دقت اندازه‌گیری سیستم‌های پردازش زبان طبیعی تأثیر زیادی دارد. برای این منظور یکی از اولین گام‌های برداشته شده، انتخاب مجموعه مناسب از داده‌های مرتبط باهدف انتخاب شده است. به‌عنوان‌مثال در متون علمی، همان‌طور که از نام آن‌ها پیداست، استفاده از جملات با ساختار عاطفی عجیب به نظر می‌رسد. همچنین استفاده از جملات احساسی در متون روایی کاملاً طبیعی به نظر می‌رسد. از سوی دیگر اگر چه منابع مکتوب زیادی وجود دارد، اما منابع احساسی برای، پردازش زبان طبیعی به‌ویژه برای زبان فارسی بسیار کم است. در تحقیقات قبلی که در مورد زبان فارسی انجام شده است معمولاً توسط نظرات کاربران در مورد یک محصول خاصی اخبار روزانه انجام شده است. در واقع هیچ مجموعه‌داده کافی و جامعی برای ارزیابی نتایج سیستم‌ها پیشنهادی وجود ندارد. این امر منجر به اولین گام در ارائه داده‌های احساسی از مجموعه توییت‌های فارسی شد که یکی از محبوب‌ترین مجموعه‌ها در زبان فارسی است. از میان منابع فارسی موجود، مجموعه توییت‌های فارسی به دلیل تنوع در مفاد اسناد، برای برچسب‌گذاری احساسی انتخاب شد. مجموعه توییت‌های فارسی مجموعه‌ای از متون فارسی است که به‌صورت حرفه‌ای با برچسب‌های دسته‌بندی نحوی برچسب‌گذاری شده‌اند. این

مجموعه شامل شش برچسب احساسی است. این مجموعه توسط زبان‌شناسان به‌صورت دستی ساخته شده و تهیه آن زمان بسیار طول کشیده.

باتوجه‌به کلماتی که افراد برای بیان احساسات خود به طور مستقیم و غیرمستقیم استفاده می‌کنند، مفاهیم عاطفی به دو اصطلاح تحت‌اللفظی و مجازی تقسیم می‌شوند. اگر معنای آن در نگاه اول آشکار نباشد و معنای آن ضمنی باشد این بنا را اصطلاح مجازی می‌نامند.

بحث انتزاع بودن احساس با گفتار مجازی بیان می‌شود. یکی از ویژگی‌های انتزاعی گفتار مجازی، استفاده مکرر از استعاره است. مثلاً احساس خشم در جمله از عصبانیت منفجر شد منفجرشدن یک انتزاع از عصبانیت در نظر گرفته شده است. استعاره‌های مفهومی این دو حوزه را از نظر برخی ویژگی‌های مشترکی که دارند به هم مرتبط می‌کنند. یک ناحیه معمولاً عینی‌تر و قابل‌لمس‌تر از دیگری است. علاوه بر استعاره‌های عاطفی، از مفاهیم کنایه نیز برای بیان احساسات استفاده می‌شود. در این حالت بر خلاف استعاره‌ها به دو حوزه مربوط نمی‌شود و فقط یک حوزه را نشان می‌دهد. هدف آن فراهم‌کردن دسترسی ذهنی به یک دامنه خاص از طریق بخشی از همان دامنه است یا سعی می‌کند با استفاده از بخشی از یک دامنه خاص به بخش دیگری از همان دامنه ارتباط برقرار کند. به‌عنوان‌مثال، عصبانیت یکی از ویژگی‌های افزایش دمای بدن است، بنابراین زمانی که فرد قرمز می‌شود، نشان‌دهنده ساختار چهره وضعیت فیزیکی فرد است.

باتوجه‌به آنچه در علوم‌شناختی در مورد عاطفه و چگونگی بازنمایی احساس توسط زبان شناخته شده است، با فرض وجود رابطه بین عاطفه و زبان نفرت، می‌توان دیدگاه‌های متعددی را مورد بررسی قرارداد.

دیدگاه اول آن است که کلمات می‌توانند احساسات درونی افراد را نشان دهند.

دیدگاه دوم نحوه بیان اصطلاحات عاطفی در سطوح آوایی، واژگانی، نحوی، معنایی و بافتی زبان است.

دیدگاه سوم آن است که سازه‌های احساسات شخصی را نشان می‌دهند؛ بنابراین ویژگی‌های مورداستفاده در سیستم پیشنهادی به شرح زیر است:

- کلمات کلیدی احساسی
- مقولات نحوی عاطفی
- ترکیب‌های احساسی

هنگام استفاده از زبان، انسان‌ها سعی می‌کنند محتوای ذهنی پیچیده خود را در قالب عینی‌تر و قابل‌دسترس بیان کنند. اگر بین عاطفه و زبان رابطه مستقیم یا غیرمستقیم وجود دارد، این رابطه چگونه نمایش داده می‌شود؟ این رابطه را به‌خوبی می‌توان به این صورت خلاصه کرد:

کلماتی مانند "او"،" هورا"و "وای" در متن به ترتیب نشان‌دهنده غم، شادی و تعجب هستند.

این واژه‌ها را واژه‌های بیانی می‌گویند. این کلمات بیانگر معنای احساسات در درون خود هستند.

کلمات کلیدی و احساسی افراد بسته به احساسی که دارند، در موقعیت‌های مختلف از کلمات متفاوتی استفاده می‌کنند. به‌عنوان‌مثال، کلماتی که افراد هنگام عصبانیت استفاده می‌کنند با کلماتی که در هنگام خوشحالی استفاده می‌کنند متفاوت است.

در این قسمت واژه‌های بیانی و تحت‌اللفظی مورد تجزیه‌وتحلیل قرار می‌گیرد؛ بنابراین فهرستی از کلمات کلیدی معنایی مرتبط با هر احساس می‌توان یافت.

بر اساس منطق انسان، واژه‌هایی که بیشترین ارتباط را با متون احساسی داشته و فراوانی بالایی در متون دارند را مرتبط می‌دانیم. سپس کلیدواژه‌های تهیه شده مورد بررسی قرار گرفت که به کدام دسته از احساسات تعلق دارند. برخی از برترین کلمات کلیدی مرتبط با احساسات در زبان فارسی هستند.

این مجموعه کلماتی که احساسات را توصیف می‌کنند، کلمات توصیفی نامیده می‌شوند. کلماتی مانند" خشم"،" شادی"،" شگفتی" کلمات توصیفی هستند. در بین کلمات توصیفی، همه کلمات به یک اندازه اهمیت ندارند و برخی از آنها نمونه بهتری هستند و نمونه اولیه احساسات به‌حساب می‌آیند. این کلمات به‌عنوان کلمات اصلی شناخته می‌شوند و گروهی دیگر واژه‌های غیر پایه نامیده می‌شوند.

در این کار پژوهشی سعی شده است از ویژگی‌های متن‌کاوی برای دستیابی به‌دقت بالا استفاده شود. باتوجه‌به ساختارهای احساسی و کلیدواژه‌ها در متون فارسی، جنبه بدیع این پروژه است. امید است شناخت کلیدواژه‌های عاطفی کارایی مناسبی کسب کند.

با توجه موارد فوق، هدف این کار ارائه سیستمی برای شناخت عاطفه بر اساس ویژگی‌های متن‌کاوی بود تا بتواند عملکردی مشابه عملکرد ذهن انسان داشته باشد. همچنین نشان می‌دهد که در عصر مدرن، استفاده از علم آمار در مطالعه قسمت‌های زیرین زبان بسیار کمک‌کننده است.

ساختار بقیه مقاله به شرح زیر است:

در فصل بعدی به بررسی معیارهای متن‌کاوی و بازیابی اطلاعات می‌پردازیم.

در فصل آخر، سیستم پیشنهادی باتوجه‌به ویژگی‌های آن شرح داده شده است. در این بخش به بررسی ویژگی‌های تعریف شده، نحوه تهیه و استفاده از آنها می‌پردازد و به میزان تأثیر و ویژگی‌های تعریف شده بر شناسایی احساسات مختلف می‌پردازد.

توجه به کلیدواژه‌ها از جمله مواردی بود که در این کار بررسی شد.

فصل ۳
مفاهیم پایه

۳-۱ متن‌کاوی

متن‌کاوی یک فرایند تحلیلی است که با استفاده از الگوریتم‌ها و تکنیک‌های مختلف، اطلاعات مخفی یا الگوهای پنهان در داده‌ها را شناسایی می‌کند. این فرایند، معمولاً به‌منظور استخراج اطلاعات مفید و نهایی از داده‌های حجیم و پیچیده انجام می‌شود.

در مرحله ابتدایی متن‌کاوی، داده‌های موردنظر از منابع مختلف جمع‌آوری می‌شوند. سپس با استفاده از الگوریتم‌های مختلف، داده‌ها تحلیل و بررسی می‌شوند تا الگوها، روابط، و اطلاعات مهم شناسایی شوند. این اطلاعات ممکن است از نظرات مشتریان، رفتارهای بازار، یا دیگر جنبه‌های مهم کسب‌وکار باشند.

یکی از روش‌های متداول در متن‌کاوی، تحلیل متنوعی از داده‌های متنی است. با بهره‌گیری از تکنیک‌های پردازش زبان طبیعی، متن‌های بلند مانند نظرات مشتریان، مقالات، یا ارسال‌های رسانه‌های اجتماعی مورد تجزیه‌وتحلیل قرار می‌گیرند. این تجزیه‌وتحلیل متن می‌تواند به شناخت احساسات، موضوعات محبوب، یا حتی شناسایی احتمالی مشکلات و نقاط قوت در محتواها منجر شود[3].

در نهایت، نتایج به‌دست‌آمده از متن‌کاوی، به تصمیم‌گیری‌های مؤثر در کسب‌وکار کمک می‌کند. این اطلاعات می‌توانند در بهبود تجربه مشتریان، افزایش کارایی عملیات، و حتی شناسایی فرصت‌های جدید در بازار کمک کنند.

در نتیجه، متن‌کاوی به‌عنوان یک ابزار قدرتمند در دنیای اطلاعاتی امروزی، برای سازمان‌ها و تصمیم‌گیرندگان ارزش زیادی ایجاد می‌کند. این فرایند تحلیلی نه‌تنها امکان شناسایی الگوها و اطلاعات مهم را فراهم می‌کند؛ بلکه به بهبود استراتژی‌ها و عملکرد کلان کسب‌وکار نیز کمک می‌کند.

۲-۳ پیش‌پردازش

در پیش‌پردازش متن چالش‌های متعددی برای زبان فارسی وجود دارد. یکی از چالش‌های سیستم‌های پردازش زبان فارسی، پیچیدگی سیستم نوشتاری آن است. به‌عنوان‌مثال، استفاده نامناسب از فاصله به‌جای نیم‌فاصله قابل‌ذکر است. توجه به استفاده از علائم نگارشی فارسی و استفاده از حروف غیرعربی می‌تواند از پیچیدگی سیستم نوشتاری بکاهد.

دیگر چالش‌ها شامل کلماتی است که می‌توان به روش‌های مختلفی نوشت. شناسایی واژه‌های مرکب و تک‌واژه‌ها درنظرگرفتن آنها از دیگر چالش‌هایی است که برای نگارش زبان فارسی وجود دارد. به‌عنوان‌مثال، پیداکردن ریشه کلمات برای درنظرگرفتن یک واحد برای کلمات مربوطه است؛ بنابراین به‌راحتی بر عملکرد سیستم تأثیر می‌گذارد.

کلمات توقف شامل کلمات دستوری مانند حروف ربط و حروف اضافه است. این کلمات فاقد اطلاعات معنایی و متنی هستند. در مورد مطالعاتی که بر فرکانس کلمه تمرکز دارند، فراوانی بالای کلمات توقف که در اسناد وجود دارد باعث خطا در نتایج پردازش می‌شود.

۳-۳ کلمات توقف

در پردازش متن، کلمات توقف به کلماتی اطلاق می‌شـود که در تحلیل و اسـتخراج اطلاعات از متن، ارزش زیادی ندارند و بیشـتر برای ارتقای کارایی و کاهش حجم داده‌ها مورداسـتفاده قرار می‌گیرند. این کلمات به‌صـورت متداول در زبان فارسـی نیز وجود دارند و در فرایند پردازش متن به‌منظور بهبود دقت و سـرعت تحلیل، اسـتفاده می‌شوند.

کلمات توقف معمولاً از دسته‌های کلمات متداول و پرتکرار در زبان فارسی هستند. این شامل حروف اضافه، حروف ربط، ضمایر و کلمات پرتکراری مانند "و"، "در"، "به" و "از" می‌شود. این کلمات توقف برخلاف کلمات مهم مانند اسم‌ها یا صفات، به‌تنهایی اطلاعات زیادی ارائه نمی‌دهند.

از اهمیت کلمات توقف می‌توان به موارد زیر اشاره کرد:

۱. کاهش حجم داده‌ها : حذف کلمات توقف باعث کاهش حجم داده‌ها می‌شـود و این امکان را فراهم می‌کند تا در تحلیل داده‌ها سریع‌تر عمل شود.

۲. بهبود دقت تحلیل : با حذف کلمات توقف، تمرکز بر روی کلمات اصلی و معنادارتر متن افزایش می‌یابد که منجر به بهبود دقت تحلیل متن می‌شود.

برای درک بهتر، مثالی از حذف کلمات توقف در یک جمله فارسـی را مورد بررسـی قرار دهیم: جمله اصـلی: "در یک روز زیبا، آسمان آبی بود و افرازهای کوه‌ها با برف پوشیده بودند."

جمله پس از حذف کلمات توقف: "روز زیبا، آسمان آبی، افرازهای کوه‌ها برف پوشیده."

در این مثال، کلمات توقف مانند "در" و "و" حذف شده‌اند تا متن کلی‌تر و معنادارتری به دست آید.

استفاده از کلمات توقف در پردازش متن فارسی بهبود سرعت و دقت تحلیل را تضمین می‌کند. این اقدام مؤثر در استخراج اطلاعات معنی‌دار از متن‌های حجیم است و در تحلیل دقیق‌تر و کارآمدتر داده‌ها نقش بسزایی ایفا می‌کند.

۳-۴ پاک‌سازی دیتاست

تمیزکردن دیتاست متنی یک گام اساسی در پردازش داده است و می‌تواند تأثیر زیادی در کیفیت و قابلیت استفاده از داده‌ها داشته باشد. در ادامه، روش‌ها و مراحلی برای تمیزکردن دیتاست متنی را بررسی می‌کنیم:

۱.حذف سطرهای خالی : حذف ردیف‌هایی که هیچ اطلاعاتی ندارند از جمله اولین گام‌هاست. این سطرها ممکن است اطلاعات ناقص یا بدون ارزشی داشته باشند.

۲. پر کردن مقادیر ناقص : اگر بخشی از داده‌ها ناقص باشد، می‌توانید از روش‌های پر کردن مقادیر ناقص مانند میانگین، مد، یا مقدار پیش‌فرض استفاده کنید. یا می توانید در صورت برخورد با چنین نوع رکوردی کل آن رکورد را حذف کنید.

۳. حذف داده‌های تکراری: اگر در دیتاست داده‌های تکراری وجود داشته باشد، حذف آنها می‌تواند به کاهش انحراف و بهبود دقت در تحلیل کمک کند.

۴. تشخیص و حذف داده‌های نویزی : داده‌های نویزی ممکن است تحلیل دقیق را مختلف کنند. با شناسایی و حذف داده‌های نویزی می‌توانید کیفیت تحلیل خود را افزایش دهید.

۵. حذف اطلاعات غیرضروری : اگر بخشی از داده‌ها برای هدف خاصی که در دست استفاده می‌شود ارزشی ندارد، آن بخش را حذف کنید.

۶. تبدیل متن به حالت نرمال: برخی از پیش‌پردازش‌های متنی مانند تبدیل به حالت نرمال، تبدیل کلمات به حروف کوچک(در نوشتار انگلیسی)، و حذف علائم نگارشی اضافی مانند ضمه، فتحه،کسره و ... می‌توانند بهبودی در تحلیل متن داشته باشند.

۷. تشـــخیص و حذف داده‌های پرتکرار : داده‌های پرتکرار ممکن اســـت موثریت تحلیل را کاهش دهند. با تشخیص و حذف این داده‌ها، دقت تحلیل افزایش می‌یابد.

۸. استفاده از کلمات توقف: حذف کلمات توقف که در توضیح قبلی آمده، به بهبود دقت تحلیل کلمات کمک می‌کند.

۹. تصـحیح املایی : اگر در داده‌ها املای اشـتباه وجود دارد، از ابزارهای تصـحیح املایی اسـتفاده کنید تا دقت اطلاعات افزایش یابد.

۱۰. تبدیل داده‌های متنی به بردارهای عددی : برای اســـتفاده از داده‌های متنی در مدل‌های عددی، می‌توانید از روش‌های تبدیل به بردارهای عددی استفاده کنید.

تمیزکردن دیتاســـت متنی مرحله مهمی در اســـتفاده مؤثر از داده‌ها اســـت. هر گاه که با دیتاســـت متنی روبرو می‌شوید، می‌توانید از این روش‌ها و تکنیک‌ها برای بهبود کیفیت و قابلیت تحلیل داده‌ها بهره‌مند شوید.

۳-۵ بازیابی اطلاعات

بازیابی اطلاعات یک حوزه مهم در علوم کامپیوتر و علم اطلاعات است که به بررسی و تحقیق در مورد روش‌ها و تکنیک‌های بازیابی اطلاعات اختصـــاص دارد. در این حوزه، هدف اصـــلی یافتن اســـناد یا اطلاعات مناســـب در مجموعه‌های بزرگ‌داده‌ها با استفاده از یک سیستم کامپیوتری است

۳-۶ مدل‌های بازیابی اطلاعات

۱. مدل محتوا [1]: در این روش، بازیابی اطلاعات بر اساس محتوای اسناد صورت می‌گیرد. ویژگی‌های مختلف مانند کلمات کلیدی، عناصر مشخص، یا ویژگی‌های محتوایی دیگر برای تطابق و جستجوی اطلاعات استفاده می‌شوند.

۲. مدل مبتنی بر کاربر [2]: این روش به اطلاعات مربوط به کاربران توجه دارد. سابقه جستجوها، نظرات یا فعالیت‌های کاربر در بازیابی اطلاعات تأثیرگذار هستند. سیستم‌های توصیه نیز معمولاً از این نوع مدل استفاده می‌کنند.

۳. مدل ترکیبی [3]: در این روش، از ترکیب مدل‌های محتوایی و مبتنی بر کاربر برای بهبود دقت و کارایی در بازیابی اطلاعات استفاده می‌شود.

۴. مدل مبتنی بر ساختار [4]: این روش در بازیابی اطلاعات از اطلاعات ساختاری و ارتباطات میان اسناد استفاده می‌کند. به عنوان مثال، اسناد متنی با ساختارهای مختلف مانند وب‌سایت‌ها یا مقالات علمی تفاوت دارند.

۵. مدل مبتنی بر احتمال [5]: این روش از مدل‌های احتمالاتی برای تعیین احتمال اطلاعات مورد جستجو استفاده می‌کند. مدل‌های مانند مدل فضای برداری احتمالاتی [6] در این دسته قرار می‌گیرند.

۶. مدل مبتنی بر هوش مصنوعی [7]: با پیشرفت هوش مصنوعی، مدل‌های مبتنی بر الگوریتم‌ها و شبکه‌های عصبی عمیق نیز در بازیابی اطلاعات مورد استفاده قرار می‌گیرند.

---

[1] Content-Based Model
[2] User-Based Model
[3] Combination Model
[4] Structure-Based Model
[5] Probabilistic Model
[6] Probabilistic Vector Space Model
[7] AI-Based Model

۷. مدل مبتنی بر ویژگی‌های مکانی [1]: در مواردی که اطلاعات مرتبط با مکان مهم هستند، مدل‌های مبتنی بر موقعیت جغرافیایی و ویژگی‌های مکانی به کار می‌روند.

هر یک از این روش‌ها و مدل‌ها بر اساس نیازها و موارد خاص مسائل بازیابی اطلاعات مورداستفاده قرار می‌گیرند. انتخاب مناسب‌ترین روش بستگی به خصوصیات داده‌ها و هدف نهایی بازیابی اطلاعات دارد.

۷-۳ مفاهیم پایه بازیابی اطلاعات

همان‌طور که در توضیحات قبلی گفته شد بازیابی اطلاعات یک حوزه گسترده در علوم کامپیوتر و علم اطلاعات است که به توسعه و بهینه‌سازی فرایند جستجو و بازیابی داده‌ها از مجموعه‌های حجیم اطلاعاتی می‌پردازد. هدف اصلی این حوزه، یافتن اسناد یا اطلاعات مرتبط با یک پرسش خاص یا یک نیاز مشخص است که در ادامه به بررسی مفاهیم اساسی و پایه این حوزه می‌پردازیم.

۱. پرس و جو [2]: در بازیابی اطلاعات، کلمات یا عباراتی که توسط کاربر به منظور جستجو از سیستم وارد می‌شوند، به عنوان کلمه کلیدی شناخته می‌شوند.

۲. سند [3]: هرگونه اطلاعات یا محتوا که ممکن است به عنوان پاسخ به یک جستجو بازیابی شود، به عنوان سند معرفی می‌شود.

۳. اطلاعات مرتبط [4]: در زمینه بازیابی اطلاعات، اطلاعاتی که با توجه به کلمه کلیدی وارد شده توسط کاربر، بیشترین تطابق را دارند و به میزان مرتبطی با جستجو، به عنوان اطلاعات مرتبط شناخته می‌شوند.

---

[1] Location-Based Model
[2] Query
[3] Document
[4] Relevance

۳-۸ روش‌ها و مراحل بازیابی اطلاعات

۱. پرسمان‌سازی[1] : در این مرحله، کاربر یا سیستم جستجوی متنی و یا تصویری خود را با استفاده از کلمات کلیدی مناسب مشخص می‌کند.

۲. جستجو و اندازه‌گیری تطابق [2]: سیستم با استفاده از الگوریتم‌های بازیابی، اسنادی که بیشترین تطابق با پرسمان دارند را انتخاب می‌کند.

۳. مرتب‌سازی نتایج [3]: اطلاعات انتخاب شده بر اساس معیارهایی نظیر تطابق، اهمیت و احتمال مرتب می‌شوند تا نتایج به کاربر به ترتیب مرتب سپرده شوند.

۴. نمایش نتایج [4]: نتایج به کاربر نمایش داده می‌شود. این ممکن است شامل متن، تصاویر، یا سایر فرمت‌های متناسب با نوع اطلاعات باشد.

۳-۹ کاربردها و دامنه‌های استفاده

۱. موتورهای جستجوی وب : موتورهای جستجو مانند گوگل از تکنیک‌های بازیابی اطلاعات برای ارائه نتایج دقیق و مرتبط به کاربران استفاده می‌کنند.

۲. سیستم‌های مدیریت اسناد : در سازمان‌ها و شرکت‌ها برای مدیریت و بازیابی اطلاعات از داکیومنت‌ها و اسناد به وسیله سیستم‌های مدیریت اسناد [5]از تکنیک‌های بازیابی استفاده می‌شود.

---

[1] Query Formulation
[2] Search and Matching
[3] Ranking Results
[4] Presentation of Results
[5] DMS

۳. پرتال‌های خبری و اطلاعاتی : سیستم‌های بازیابی اطلاعات در پرتال‌های اخباری جهت ارائه اطلاعات مرتبط با جستجوهای کاربران به کار می‌روند.

۴. پرسش و پاسخ هوشمند : در سامانه‌های پرسش و پاسخ هوشمند، از تکنیک‌های بازیابی اطلاعات برای یافتن پاسخ‌های مرتبط با سوالات کاربران استفاده می‌شود.

بازیابی اطلاعات به‌عنوان یک حوزه اساسی در علوم کامپیوتر، در جوانب مختلف زندگی و تکنولوژی تأثیرگذار بوده و با پیشرفت تکنولوژی، روش‌ها و مدل‌های جدیدی برای بهبود عملکرد در این زمینه توسعه می‌یابد.

۳-۱۰ برخی از تکنیک‌های بازیابی اطلاعات

تکنیک‌های بازیابی اطلاعات متنوع و گسترده هستند و باتوجه‌به نوع اطلاعات موردنظر، میزان دقت موردنیاز، و ساختار داده‌ها، از روش‌های مختلف استفاده می‌شوند. در زیر به برخی از تکنیک‌های معروف بازیابی اطلاعات اشاره می‌شود:

۱. مدل فضای برداری [1]: این مدل اطلاعات را در یک فضای چند بعدی نمایش می‌دهد و از ابزارهای ریاضی برای اندازه‌گیری تطابق میان پرسمان و اسناد استفاده می‌کند.

۲. مدل BM۲۵ : این مدل مبتنی بر احتمالات است و برای اندازه‌گیری تطابق بین پرسمان و اسناد از توزیع احتمالاتی BM۲۵ استفاده می‌کند.

[1] Vector Space Model

۳. پردازش زبان طبیعی [1]و تحلیل متن : اســتفاده از تکنیک‌های پردازش زبان طبیعی مانند تحلیل موضــوع، خوشه‌بندی متن، تحلیل احساسات، و استخراج اجزاء مهم متن برای افزایش دقت در بازیابی اطلاعات متنی.

۴. شــبکه‌های عصــبی : از شــبکه‌های عصــبی عمیق برای یادگیری ارتباطات پیچیده بین کلمات و ســاختارهای مختلف در متن‌ها جهت بهبود بازیابی اطلاعات استفاده می‌شود.

۵. شـباهت کسـینوسـی [2]: این تکنیک بر اسـاس زاویه میان بردارهای متنوع در فضـای بردارهای عددی اسـتفاده می‌شود و میزان شباهت بین پرسمان و اسناد را اندازه‌گیری می‌کند.

۶. الگوریتم‌های هش [3]: برای شــتاب‌زدن فرآیند بازیابی اطلاعات، الگوریتم‌های هش برای نمایش ســریع و بهینه‌تر داده‌ها مورد استفاده قرار می‌گیرند.

۷. الگوریتم‌های درخت تصـمیم[4] : این الگوریتم‌ها به منظور کلاس‌بندی و جسـتجوی بهینه در فضـای ویژگی‌ها برای افزایش دقت و سرعت در بازیابی اطلاعات مورد استفاده قرار می‌گیرند.

۸. مدل‌های ترکیبی [5]: این مدل‌ها از ترکیب چندین مدل مختلف با هدف بهبود دقت و کارایی در بازیابی اطلاعات استفاده می‌کنند.

هرکدام از این تکنیک‌ها و روش‌ها به‌مرورزمان با پیشــرفت تکنولوژی و علوم داده توســعه می‌یابند و برای حل چالش‌های مختلف در بازیابی اطلاعات مورداستفاده قرار می‌گیرند.

[1] NLP
[2] Cosine Similarity
[3] Hashing Algorithms
[4] Decision Trees
[5] Ensemble Models

۳-۱۱ معیارهای امتیازدهی

معیارهای امتیازدهی در بازیابی اطلاعات جهت ارتباط دقیق‌تر با نیازهای کاربران مورداستفاده قرار می‌گیرند. این معیارها عموماً بر اســـاس تحلیل تکرار و اهمیت واژگان در متن‌ها شـــکل می‌گیرند. امتیازدهی به‌عنوان یک فرایند انتخابی، با بهره‌گیری از معیارهای مختلف مانند تعداد تکرار[1]، تعداد اســـناد حاوی یک واژه [2]، اهمیت مفهومی یک واژه در متن [3]و ترکیبی از انها یعنی TF-IDF، به ما کمک می‌کند تا اســـناد مرتبط را بادقت بالاتری انتخاب کنیم. این معیارها اساسی‌ترین ابزارها برای بهبود عملکرد سامانه‌های بازیابی اطلاعات هستند، زیرا امکان تفکیک بین اطلاعات مرتبط و غیرمرتبط را فراهم می‌کنند.

در ادامه به بررسی دقیق‌تر معیارها می‌پردازیم.

۳-۱۱-۱ تی اف

در حوزه بازیابی اطلاعات، معیار TF[4] یکی از مفاهیم کلیدی اســت که فرکانس تکرار یک کلمه خاص در یک ســند را اندازه‌گیری می‌کند. این معیار نشان‌دهنده میزان اهمیت یک کلمه در یک ســند خاص اســت در ســیســتم‌های بازیابی اطلاعات، از معیار TF برای رتبه‌بندی اسناد بر اساس میزان تطابق با کلمات کلیدی پرسمان استفاده می‌شود، درعین‌حال این معیار به‌تنهایی نمی‌تواند رتبه خوبی ارائه کند. در پردازش زبان طبیعی، معیار TF به‌عنوان یک ویژگی مهم برای تحلیل و اســتخراج ویژگی‌های متنی مورداســتفاده قرار می‌گیرد. در مســائل دســته‌بندی متون، از معیار TF به‌عنوان یکی از ویژگی‌های ورودی برای مدل‌های یادگیری ماشین استفاده می‌شود. معیار فرکانس تعداد تکرار کلمه TF یک ابزار در تحلیل محتوای متنی اســت که با محاســبه تعداد تکرار یک کلمه در یک ســند، اطلاعات مهمی در

[1] TF
[2] DF
[3] IDF
[4] Term Frequency

مورد اهمیت و تأثیر آن کلمه در سند ارائه می‌دهد. این معیار یک روش برای اندازه‌گیری اهمیت یک کلمه در یک سند نسبت به یک مجموعه یا کورپوس از اسناد است. این مقدار می‌تواند به‌صورت تعداد خام، یا به‌صورت نرمال شده با تقسیم بر تعداد کل کلمات در سند یا به‌صورت لگاریتمی یا به‌صورت نرمال دوگانه، محاسبه شود.

۳-۱۱-۱-۱ رابطه ریاضی

TF بیان می‌کند که i نشان‌دهنده کلمه موردنظر و j نمایانگر سند است. این معیار به‌صورت زیر محاسبه می‌شود

$$TF=\frac{\text{تعداد تکرار کلمه در سند}}{\text{تعداد کل کلمات در سند}}$$

۳-۱۱-۲ دی اف

معیار DF [1] یکی دیگر از معیارها در حوزه بازیابی اطلاعات است که میزان پراکندگی یا فراوانی یک کلمه در میان اسناد مختلف را اندازه‌گیری می‌کند. این معیار برای تشخیص اهمیت یک کلمه در مجموعه اسناد مورداستفاده قرار می‌گیرد. این معیار نشان‌دهنده تعداد اسنادی است که حاوی یک کلمه خاص هستند و به‌عبارت‌دیگر، میزان پراکندگی یا روند فراوانی آن کلمه در کل مجموعه اسناد. از DF به‌عنوان یکی از معیارهای انتخاب کلمات کلیدی استفاده می‌شود. کلمات کلیدی با DF کمتر ممکن است اطلاعات خاصی را نمایان کنند. در تحلیل و پردازش متن، استفاده از DF به‌عنوان ویژگی می‌تواند کمک کند تا کلماتی که در اسناد مختلف پراکنده‌تر هستند را شناسایی و

[1] Document Frequency

تحلیل کرد. معیار فراوانی سند ابزاری قدرتمند در تحلیل و بازیابی اطلاعات متنی است. با بهره‌وری استفاده از این معیار، می‌توان نتایج دقیق‌تر و مفیدتری در پروژه‌های مختلف دریافت کرد.

۳-۱۱-۲-۱ رابطه ریاضی

DF برابر با تعداد اسناد حاوی کلمه i است.

$$DF = \text{تعداد اسناد حاوی کلمه}$$

۳-۱۱-۳ مقایسه DF و TF

۱. استفاده در بازیابی اطلاعات

TF: برای اندازه‌گیری اهمیت یک کلمه در یک سند خاص و مرتبط با پرسمان‌ها.

DF: برای انتخاب کلمات کلیدی که در مجموعه اسناد به‌عنوان مهم شناخته شده‌اند.

۲. تاثیر تعداد کلمات

TF: توجه به نسبت تکرار به تعداد کلمات در یک سند.

DF : مستقل از تعداد کلمات در یک سند و فقط به تعداد اسناد حاوی کلمه متمرکز است.

۳. کاربرد در مسائل مختلف

TF : برای مسائل مربوط به تحلیل محتوا و استخراج ویژگی‌ها.

DF : برای مسائل مربوط به انتخاب و تشخیص کلمات کلیدی و اهمیت آنها در میان اسناد.

درکل، هرکدام از معیارهای TF و DF وظایف خاص خود را دارند و در مسائل متفاوت مورداستفاده قرار می‌گیرند. TF بر روی هر سند تمرکز دارد، درحالی‌که DF بر روی کلمات موجود در کل مجموعه اسناد تمرکز دارد. در مسائل بازیابی اطلاعات، معمولاً از ترکیب این دو معیار برای بهترین نتایج استفاده می‌شود.

۳-۱۱-۴ ای دی اف

معیار IDF [1]یکی از ابزارهای حوزه بازیابی اطلاعات و پردازش متن است. این معیار اهمیت یک کلمه خاص در یک مجموعه اسناد را اندازه‌گیری می‌کند. محاسبه این معیار به این اعتقاد بنیان‌گذاری شده است که کلمات که در تعداد کمی از اسناد مجموعه ظاهر می‌شوند، اهمیت بیشتری دارند.IDF نشان‌دهنده عکس میزان تکرار یک کلمه در اسناد مجموعه است. با افزایش تعداد اسنادی که حاوی کلمه موردنظر هستند، ارزش IDF کم می‌شود. معیار IDF برای تعیین اهمیت کلمات کلیدی در یک پرونده یا پرسمان استفاده می‌شود. در مسائل تحلیل متن، معیار IDF برای شناسایی کلمات کلیدی و ویژگی‌های مهم در متون به کار می‌رود. این مقدار با تقسیم تعداد کل اسناد، N، بر تعداد اسنادی که کلمه t در آن‌ها ظاهر شده است، df(t)، به دست می‌آید. idf(t) = N / df(t). این مقدار نشان می‌دهد که کلمات کمتر رایج اهمیت بیشتری دارند. برای جلوگیری از تأثیر زیاد اعداد بزرگ، معمولاً از لگاریتم این مقدار استفاده می‌شود.

[1] Inverse Document Frequency

۳-۱۱-۴-۱ رابطه ریاضی

معیار IDF برای یک کلمه از رابطه زیر به دست می‌آید:

$$IDF = \log(\frac{\text{تعداد کل اسناد}}{\text{تعداد اسناد حاوی کلمه} + 1})$$

۳-۱۱-۴-۲ مقایسه با معیارهای قبلی

تفاوت: DF تعداد اسناد حاوی کلمه را نشان می‌دهد، TF معیار تکرار کلمه در یک سند است، درحالی‌که IDF تأثیر فراوانی آن کلمه در کل مجموعه اسناد را نشان می‌دهد.

استفاده: DF برای انتخاب کلمات کلیدی در یک سند خاص،TF برای تحلیل محتوا و استخراج ویژگی‌ها مورداستفاده قرار می‌گیرد، درحالی‌که IDF برای انتخاب کلمات کلیدی در مقیاس گسترده‌تر مورداستفاده قرار می‌گیرد.

۳-۱۱-۵ تی اف ای دی اف

معیار TF-IDF [1] یکی از مهم‌ترین و مؤثرترین ابزارهای استفاده در بازیابی اطلاعات و پردازش متن است. این معیار اهمیت یک کلمه خاص در یک سند خاص را مشخص می‌کند و به‌عنوان یک ویژگی مؤثر در تحلیل و بازیابی اطلاعات متنی به کار می‌رود. معیار TF-IDF نشان‌دهنده ترکیب تأثیر ترم فراوانی و معکوس فراوانی اسناد بر یک کلمه خاص است. این معیار نمایانگر اهمیت یک کلمه در یک سند خاص و ارتباط آن با سایر اسناد مجموعه است. معیار TF-IDF به‌عنوان یک ابزار کامل و مؤثر در تحلیل متن و بازیابی اطلاعات در پروژه‌های مختلف مورداستفاده

[1] Inverse Document Frequency - Term Frequency

قرار می‌گیرد. استفاده صحیح از این معیار می‌تواند به‌دقت و اثربخشی سیستم‌های پردازش متن و بازیابی اطلاعات کمک کند. TF-IDF به‌عنوان یکی از ویژگی‌های کلیدی در مدل‌های بازیابی اطلاعات مورداستفاده قرار می‌گیرد. این معیار میزان مطابقت میان پرسمان و اسناد را بهبود می‌بخشد. در تحلیل متن و استخراج اطلاعات مفید، TF-IDF برای تشخیص کلمات کلیدی و اهمیت آنها در یک متن مورداستفاده قرار می‌گیرد. برای پیاده‌سازی، می‌توان از مقادیر ترم فراوانی و معکوس فراوانی اسناد برای هر کلمه در هر سند استفاده کرد و با ضرب دو مقدار حاصل، مقدار TF-IDF را محاسبه کرد.

در واقع هدف این سیستمِ وزن‌دهی، نشان‌دادن اهمیت کلمه در متن است که اغلب در جستجوهای درون بازیابی اطلاعات، متن‌کاوی و مدل‌سازی کاربر استفاده می‌شود. مقدار TF-IDF به‌تناسب تعداد تکرار کلمه در سند افزایش می‌یابد و توسط تعداد اسنادی که در مجموعه هستند و شامل کلمه نیز هستند متعادل می‌شود. به این معنی که اگر کلمه‌ای در بسیاری از متون ظاهر شود احتمالاً کلمه‌ای متداول است و ارزش چندانی در ارزیابی متن ندارد. در حال حاضر TF-IDF یکی از محبوب‌ترین روش‌های وزن گذاری اصطلاحات می‌باشد و امروزه بیش از ۸۳ درصد از سامانه‌های توصیه‌گر در کتابخانه‌های دیجیتال از این روش وزن‌دهی اصطلاحات استفاده می‌کنند.

این اختلاف بین وزن‌ها که توسط روش TF-IDF ایجاد می‌شود توسط بیشتر موتورهای جستجو به‌عنوان ابزار اصلی رتبه دهی و امتیازدهی اسناد پرس‌وجو شده کاربر استفاده می‌شود؛ و همچنین برای فیلترکردن ایست‌واژه‌ها در زمینه‌های موضوعی مختلف، از جمله خلاصه‌سازی و دسته‌بندی متن با موفقیت استفاده شده است. یکی از ساده‌ترین تابع‌های رتبه‌بندی با جمع‌کردن وزن به‌دست‌آمده توسط TF-IDF برای هر اصطلاح پرس‌وجو محاسبه می‌شود. بسیاری از توابع رتبه‌بندی پیچیده‌تر بر اساس این مدل ساده به وجود آمده‌اند.

معیار TF-IDF برای کاربردهای مختلفی در حوزهٔ بازیابی اطلاعات، داده‌کاوی متن، و مدل‌سازی کاربر مفید است. برخی از این کاربردها عبارت‌اند از:

رتبه‌بندی و ارزیابی اسناد: با استفاده از معیار TF-IDF می‌توان اسناد را بر اساس اهمیت کلماتی که در آن‌ها ظاهر شده‌اند، رتبه‌بندی و ارزیابی کرد. این روش به موتورهای جستجو کمک می‌کند تا اسناد مرتبط با پرسمان کاربر را پیدا و نمایش دهند.

خلاصه‌سازی متن: با استفاده از معیار TF-IDF می‌توان کلمات کلیدی یک متن را شناسایی کرد و با استفاده از آن‌ها یک خلاصه از متن ایجاد کرد. این روش به سیستم‌های خلاصه‌سازی متن کمک می‌کند تا محتوای اصلی متن را با حفظ اهمیت کلمات، بازنمایی کنند.

توصیه‌دهی متن: با استفاده از معیار TF-IDF می‌توان شباهت بین اسناد را محاسبه کرد و بر اساس آن اسناد مشابه به کاربر توصیه کرد. این روش به سیستم‌های توصیه‌دهی متن کمک می‌کند تا اسنادی را که کلمات مشترک و مهم با سند موردعلاقه‌ی کاربر دارند، پیشنهاد دهند

۳-۱۱-۵-۱ رابطه ریاضی

معیار TF-IDF برای یک کلمه در یک سند خاص به‌ازای مجموعه اسناد به‌صورت زیر محاسبه می‌شود:

$$TF - IDF(t.d.D) = TF(t.d) * IDF(t.D)$$

۳-۱۱-۵-۲ مقایسه با معیارهای قبلی

تفاوت DF : تنها تعداد اسناد حاوی کلمه را نشان می‌دهد، TF نمایانگر تعداد تکرار یک کلمه در یک سند است، درحالی‌که TF-IDF ترکیبی از میزان تکرار کلمه در یک سند و اهمیت آن در کل مجموعه اسناد است.

استفاده DF : برای انتخاب کلمات کلیدی در یک سند خاص، TF بیشتر در تحلیل محتوا و استخراج ویژگی‌ها مورداستفاده قرار می‌گیرد، درحالی‌که TF-IDF برای انتخاب کلمات کلیدی باتوجه‌به ارتباطات بین کلمات در کل مجموعه و تحلیل و بازیابی اطلاعات در مقیاس گسترده‌تر مفید است

فصل ٤
مطالعات و تحقیقات

۴-۱ روند کلی

در این فصل به برسی مدل انتخاب شده می‌پردازیم با توجه به مقدمات توضیح داده شده در فصل قبلی می توانیم با استفاده از معیار های امتیاز دهی اقدام به محاسبه با استفاده از معیار انتخابی کنیم.

نکته حائض اهمیت در این اقدام این است که معیار انتخاب شده بتواند بخوبی فاصله را میان مفاهیم احساسی کلمات ایجاد کند. در فرم اولیه دیتاها هر سطر حاوی ۷ مقدار است که اولین مقدار آن متن توییت است و ۶ مقدار بعدی تشخیص یافته به آن متن بر اساس ۶ احساس می باشد.

به فرمت داده‌ها در جدول ۴-۱ توجه کنید.

جدول ٤-١ داده‌های اصلی

۱	متن	خشم	ترس	شادی	نفرت	غم	تعجب
۲	خوش باشن من که یادم رفته دیگ :)	۲	۰	۱	۱	۱	۱
۳	باشد که #کرونا یک حقیقت را در گوش سنگین و مغز زنگ زده برخی فروکند: هیچ کشوری جزیره‌ای جدا از دیگران نیست؛ اقتصاد، سلامت، امنیت، سیاست و محیط‌زیست همه به هم وابسته است.	۰	۰	۰	۲	۱	۰
۴	کرونا رو شکست میدهیم؟ مرحله بعد چه کاری می خوای بکنی؟!	۴	۳	۱	۳	۳	۴
۵	اگر در چند ماه اخیر تصمیم داشته اید وارد بورس شوید اما به هر دلیلی به تاخیر افتاده است، این چند روز را هم صبر کنید تا تکلیف مشخص شود. شاید لازم باشد ورود به بازار سرمایه را کلا فراموش کنید. اگر برگشتی در کار بود کمی بالاتر ورود کنید. شاید آقای مددی همه واقعیت را نگوید.	۲	۴	۱	۲	۴	۲
۶	احساس می‌کنم اگر همین الان نخوابم، تمام مواد شیرینی که دارم توی خونه رو می‌خورم	۱	۱	۰	۰	۳	۲
۷	اینقدر گرفتار مسایل میشی که تخصص از دستت درد میره و به این فکر میکنی چطوری زنده بمونم	۲	۱	۰	۱	۰	۴
۸	ببینید این جلد های رنگارنگتان را که روی هیولاهای درونتان ساخته اید"	۱	۱	۰	۱	۱	۱

۲-۴ پیش‌پردازش

۱. باتوجه‌به اینکه توییت‌های ما حاوی جملات زبان فارسی است و هدف ما در انتها تشخیص کلمات تأثیرگذارتر در هر حس است، بنابراین بسیاری از کلمات در کلیه رشته‌ها تکراری است و فاقد ارزش است؛ بنابراین کلیه کلماتی که در سه فایل از پیش آماده‌شدهٔ را برای حذف از توییت‌ها مدنظر داریم.

جدول ٤-٢ فایل‌های پیش پردازش

۱	کوتاه	فعلی	غیرفعل
۲	و	کن	دیگران
۳	در	کرد	همچنان
۴	به	کردن	مدت
۵	از	باش	چیز
۶	که	بود	سایر
۷	می	بودن	جا
۸	این	شو	طی
۹	است	شد	کل
۱۰	را	شدن	کنونی
۱۱	با	دار	بیرون
۱۲	های	داشت	مثلا
۱۳	برای	داشتن	کامل
۱۴	آن	خواه	کاملا
۱۵	یک	خواست	آنکه
۱۶	شود	خواستن	موارد
۱۷	شده	گوی	واقعی
۱۸	خود	گفت	امور
۱۹	ها	گفتن	امورات

۲. حذف کردن کلیه کلمات غیر از الفبای فارسی

۳. حذف ستون نام احساسات که در اولین سطر داده های اصلی موجود است

۴. تعیین رفتار با مقدار وارد نشده مانند رکورد های خالی ، حذف رکورد هایی اعضای آن برابر با ۷ نباشد

۵. تبدیل دیتاها به لیست در فرمت جدید

۶. چک مجدد اینکه آیا پس از اعمال پیش پردازش ها متن رکورد خالی شده باشد

۷. بازگرداندن دیتاها به برنامه

۴-۳ رویکرد ما در عملیات بایس کردن استاپ وردها

سطر مربوط به هر داده‌ها را با معیار جداسازی (" ") از هم جدا می‌کنیم و به یک لیست جدید انتقال می‌دهیم و روی لیست جدید که حاوی کلمات است پیمایش انجام می‌دهیم و هر کجا به کلمه‌ای برخورد کردیم که در هریک از سه فایل پیش‌پردازش موجود باشد، آن را حذف خواهیم کرد.

۴-۴ محاسبه TF-IDF

به‌منظور محاسبه TF-IDF کلمات نیاز به یک لیست یونیک از کلمات داخل دیتاست داریم تا از محاسبه TF-IDF تکراری برای کلمات جلوگیری شود.

رویکرد ما در تشخیص لیست کلمات یونیک، تبدیل دیتاها به فرمی است که در ادامه قابل‌استفاده باشد.

با حلقه روی هر سطر از دیتاها پیش‌پردازش شده تشکیل یک لیست جدید و جاگذاری متن توییت در اولین عنصر آن، و تشکیل یک لیست موقت برای جای‌گذاری هر یک از شش احساس دیگر و در انتها قراردادن لیست موقت در دومین عنصر لیست اصلی و تکرار این عملیات که در انتها منجر به دستیابی به لیستی است که حاوی دیتاها با فرمت موردنظر ما است.

در انتها این مرحله در صورت موفقیت پیام موفق بودن عملیات چاپ می‌شود.

۴-۵ اعلان متغیر

باتوجه‌به این موضـــوع که درصـــورتی‌که TF-IDF از مقداری پایین‌تر شـــود و ورودی متنی بلند باشـــد بر روی عملکرد برنامه، تأثیر منفی می‌گذارد پس‌نیاز است یک معیاری بگذاریم و در صوت مواجه‌شدن TF-IDF های کمتر از آن آنها را در نظر نگیریم و از آن گذر کنیم.

باتوجه‌به ماهیت دیتاست و وجود امتیاز برای هر یک از احساسات شش‌گانه درصورتی‌که با متنی مواجه شدیم که امتیاز یک احساس آن زیر یک عدد خاص باشد در محاسبه TF-IDF آن کلمه در جمله‌ای که احساس مربوطه آن زیر یک مقدار خاص باشد، آن حس وTF-IDF اش را نادیده می‌گیریم. در جهت بهبود کارایی این عمل و کنترل راحت‌تر آن دو متغیر تعریف می‌کنیم و آن را در متن اصلی برنامه مقداردهی می‌کنیم.

۴-۶ مدل ذخیره‌سازی TF-IDF

روند کلی به این صـــورت اســـت که یک پیمایش روی لیســـت کلمات یونیک می‌زنیم و TF-IDF آن کلمه در لیست کلیه توییت‌ها و با توجه امتیاز آن توییت که وزن آن TF-IDF محسوب می‌شود، حساب می‌کنیم.

به طور دقیق‌تر تابع محاســـبه کننده TF-IDF ســـه مقدار کلمه، مجموعه توییت و امتیاز احســـاســـات هر جمله را دریافت می‌کند و شش متغیر به‌ازای هر یک از احساسات با مقدار اولیه ۰ اعلان می کند و پس از آن اقدام به محاسبه IDF آن کلمه در مجموع توییت ها می کند پس از آن با اعمال یک پیمایش روی کلیه توییت ها و محاســـبه tf کلمه ورودی در توییت پیمایش شده چک می کنیم که مقدار TF آن صفر نباشد تا از عملیات اضافه جلوگیری کنیم زیرا در صـــورتی که هر یک از متغیر های TF-IDF صـــفر باشـــد مقدار TF-IDF صـــفر خواهد بود . در ادامه حلقه ای که توییت ها را پیمایش می کند و مقدار TF را محاســـبه می کنیم، مقدار بدســـت آمده برای TF-IDF را که در وزن آن

توییت ضـرب شـده اسـت را به مقدار مربوط به آن احسـاس اضـافه می کنیم و در انتها یک لیسـتی حاوی شـش متغیر مقدار TF-IDF برای هریک از شش احساس است بازمی گردانیم.

مقادیر بازگردانده شـده نمایش‌دهنده امتیازهای مختلف آن کلمه در شـش احسـاس هسـتند که این مقادیر را در یک دیکشـنری با سـاختمان داده‌ای به این شـکل که کلید هر مقدار، کلمه خواهد بود و مقدار آن برابر با همان لیسـتی است که از تابع محاسبه کننده TF-IDF دریافت کرده‌ایم

. به علت حجم بالای پردازش اطلاعات دیکشـنری نهایی را ذخیره می‌کنیم تا در ادامه مسـیر و انجام تسـت روی داده از همین مدل تولید شده استفاده کنیم و مجبور نباشیم هر بار مقدار TF-IDF کلمات را مجدداً محاسبه کنیم.

تا اینجای کار ما موفق به تولید مدل شده‌ایم در ادامه می‌توانیم جزئیات مدل را مشاهده کنیم.

جدول ٤-٣ کلمه بر حسب TF-IDF برای هر شش احساس

۱	کلید	خشم	ترس	شادی	نفرت	غم	شگفتی
۲	مرحله	۱۳.۱۰	۶.۰۷	۱۱.۸۹	۹.۰۱	۸.۵۶	۷.۳۵
۳	میخوای	۱۵.۰۶	۵.۱۲	۲.۱۲	۱۲.۶۴	۷.۴۱	۱۸.۲۹
۴	بخوری	۱۳.۱۱	۵.۶۲	۰	۷.۴۲	۸.۷۶	۹.۹۶
۵	ماه	۱۳.۳۹	۱.۸۱	۷.۸۶	۱۳.۲۵	۲۷.۴۸	۱۰.۱۴
۶	تصمیم	۲.۸۶	۱.۹۲	۳.۵۰	۸.۸۳	۴.۰۸	۴.۳۷
۷	اید	۲.۸۸	۰. ۸۲	۰	۳.۲۹	۲.۹۵	۱.۸۵
۸	وارد	۱۱.۰۹	۵.۲۹	۱۰.۵۹	۸.۱۲	۶.۴۰	۱۳.۰۱
۹	بورس	۶۶.۶۴	۴۶.۰۰	۶.۳۲	۳۹.۴۷	۸۶.۱۶	۵۶.۸۸
۱۰	شوید	۲.۴۵	۰.۸۹	۰	۲.۲۳	۱۳.۸۹	۰
۱۱	دلیلی	۹.۰۹	۶.۱۳	۰.۶۳	۰	۱.۴۸	۰.۶۳
۱۲	تاخیر	۱.۹۱	۰.۹۶	۰	۰	۰.۹۶	۰
۱۳	افتاده	۰.۸۹	۱.۲۵	۳.۵۶	۱.۹۶	۷.۰۰	۱.۰۷

۴-۷ تست داده های جدید

گفته شد محاسبه جدول TF-IDF در لحظه و بصورت بلادرنگ ممکن نیست و بار محاسباتی دارد بدین منظور دیتاها را ذخیره کردیم و جهت استفاده مجدد مدل تولید شده را فقط به برنامه اضافه می کنیم و دوباره آن را تولید نمی کنیم. به جهت سهولت در کدنویسی و مشخص بودن دیتاها، شش لیست خالی تعریف می کنیم و در این لیست ها داده هایمان را به شکل یک لیست دوتایی شامل کلمه و مقدار TF-IDF آن کلمه در یک حس خواهد بود. در انتها شش لیست خواهیم داشت که در هر لیست کلیه کلمات یک حس همراه با ارزش آن کلمه در همان حس موجود می باشد

به جهت پیش بینی از دیکشنری قبلی استفاده می کنیم بدین صورت که یک حلقه روی تمام کلیدها و مقدار کلیدها می زنیم و پیمایش انجام می دهیم.

۴-۸ روش‌های تست ورودی

این قسمت از برنامه به نحوی نوشته شده است که در دو حالت تست داده های دیگر و اجرای به تنهایی می تواند استفاده شود در صورتی که این برنامه منحصرا اجرا شود اقدام به دریافت ورودی از کاربر می کند و عملیات تشخیص حس را انجام می دهد، حالت دیگر به این صورت است که در یک برنامه دیگر که دیتا های تست را آماده کرده ایم این تابع را صدا می زنیم و خروجی ما به ازای یک رشته تنها حس تشخیص داده شده خواهد بود در صورتی که در حالت اجرای انحصاری کد بدون فراخوانی در کد دیگر جزئیات بیشتری نمایش داده می‌شود.

۹-۴ نحوه تشخیص احساس از TF-IDF

رویکرد انتخابی ما در تشخیص و انتخاب حس به‌عنوان احساس غالب متن بدین شکل خواهد بود که با دریافت رشته ورودی شش متغیر برای نگهداری مقدار امتیاز متن به‌ازای هر شش حس به‌صورت جدا از هم تعریف می‌کنیم.

رشـــته وردی را با معیار قراردادن " " از هم جدا می‌کنیم و در یک لیســـت قرار می‌دهیم و پیمایش را روی آن انجام می‌دهیم.وضعیت هر کلمه در هر شش حس را حساب می‌کنیم. بدین صورت که اگر کلمه پیمایش شده در هر لیسـت تولید شـده هر حس وجود داشـته باشـد، مقدار TF-IDF آن بازگردانده می‌شـود و در غیر این صـورت مقدار منطقی نادرست[1] بازگردانده می‌شود سپس با بررسی وضعیت کلمه در هر حس درصورتی‌که مقدار مخالف نادرست باشد TF-IDF بازگردانده شـده را به متغیرهای اصلی که نگهدارنده مقدار کل TF-IDF های هر حس هستند اضافه می‌کنیم. در انتها از میان شش متغیر حسی که بیشترین مقدار عددی را دارا باشد را به‌عنوان احساس قالب متن معرفی می‌کنیم.

۱۰-۴ دقت مدل

برای محاسبه دقت مدل از یک دیتاستی استفاده می‌کنیم که حاوی متن و نوع احساس آن می‌باشد [4].

جدول ٤–٤ نمونه از داده های تست

1	کلید	مقدار
۲	خیلی کوچک هستن و سایزشون بدرد نمیخوره میخوام پس بدم	غمگین
۳	از صدای پرنده دم دمای صبح متنفرم متنفرم متنفرم	نفرت
۴	این وضع به طرز خنده داری گریه داره...	غمگین
۵	خب من رسما از یک نفر متنفرم،چون از گربه بدش میاد ازصبح شروع کرده رو مخ من راه رفتن ؛شپش میگری،کز از میگیری،کک داره،درد داره،مرض داره	نفرت

[1] False

دقت مدل تولید شده باتوجه‌به مطالب گفته شده روی دیتاهای تست به شکل زیر خواهد بود

جدول ٤–٥ کد نمونه‌ها

نمونه‌ها	پیش بینی های درست	دقت	ترس	نفرت	غمگین	خشمگین	خوشحال	تعجب
۱۲۴۰	۱۰۹۰	۸۷.۹۰۳۲۲۵	۵۵۱	۵۱	۵۹	۴۱۲	۶۸	۹۹

فصل ۵
نتیجه گیری و پیشنهادات

۵-۱ نتیجه گیری

همانطور که گفته شد هدف از این پروژه ارائه ی ساز و کاری نوین در تشخیص احساسات با استفاده از روش های بازیابی اطلاعات بوده است. آنچه اهمیت این پروژه را دوچندان کرده بود عدم وجود اقدامات مشابه در زبان فارسی بود که با وجود منابع بسیار اندک توانستیم به مقدار زیادی این پروژه را در حالت بهینه ارائه کنیم.

در انتهای این راه ما می توانیم با دریافت یک متن از ورودی احساس آن را تشخیص دهیم، این امر می تواند در زمینه های بازاریابی و فروش بسیار کابردی باشد و می توانیم آن را مسائل مختلفی استفاده کنیم.

در ادامه بیشتر به اهمیت متن کاوی در متون ادبیات فارسی اشاره می کنیم.

متن کاوی در متون ادبیات فارسی اهمیت زیادی دارد و می‌تواند به درک عمیق‌تر از ساختارها، مضامین، و احساسات مختلف متون ادبی کمک کند. در زیر توضیحاتی در مورد اهمیت این موضوع ارائه شده است:

تفکر علمی و فرهنگی : متون ادبیات فارسی غنی از تفکر علمی و فرهنگی هستند. متن کاوی در این متون به ما امکان می‌دهد تا درک عمیقی از این تفکرات را پیدا کنیم و برای شناخت بهتر فرهنگ و ادبیات ایرانیان اقدام نماییم.

شناخت احساسات و عواطف : ادبیات فارسی غنی از عواطف و احساسات انسانی است. متن کاوی در این حوزه به ما این امکان را می‌دهد تا عمیق‌ترین احساسات نویسندگان را درک کرده و با زبان ادبی انتقال دهیم.

شناسایی نقش و تأثیر نویسنده : متون ادبیات فارسی در بسیاری از موارد تحت تأثیر تجربیات و دیدگاه‌های نویسنده قرار دارند. با استفاده از متن کاوی، ما می‌توانیم نقش و تأثیر نویسنده در آثارش را بهترین طور ممکن شناسایی کنیم.

تحلیل ساختار و اندیشه : ادبیات فارسی اغلب دارای ساختارها و اندیشه‌های پیچیده‌ای هستند. متن کاوی به ما این امکان را می‌دهد تا اجزا و عناصر مختلف یک متن را تجزیه و تحلیل کنیم و به فهم بهتری از ساختار ادبی و اندیشه نویسنده برسیم.

تفسیر ساختارهای زبانی : زبان ادبی فارسی دارای ساختارهای زبانی خاصی است. متن کاوی در این حوزه به ما این امکان را می‌دهد که ساختارهای زبانی پیچیده را تجزیه و تحلیل کرده و به درک عمیق‌تری از زبان ادبی فارسی برسیم.

کمک به پژوهش‌های بیشتر : نتایج متن کاوی می‌توانند به پژوهش‌های بیشتر در زمینه ادبیات فارسی و متون مشابه کمک کنند و محققان را به سمت پرسش‌ها و چالش‌های جدید هدایت کنند.

این نکات تنها یک آغاز برای شروع بحث در مورد اهمیت متن کاوی در متون ادبیات فارسی هستند.حال که جزئیات بیشتری از اهمیت این موضوع می دانیم می توانیم بهتر در مورد بررسی تحقیقات انجام شده اظهار نظر نمود.

در تحقیقات این پروژه ما ابتدا سعی کردیم دیتاست را بر اساس اصول اولیه نرمال کردن آنرا به یک حداقل برای اجرا کردن روش های امتیاز دهی تبدیل کنیم. در این مسیر ما می توانیم از گفایل های اماده که برای مثال کلمات توقف را گسترش داده اند کمک بگیریم اما با توجه به ماهیت فایل های ما در این پروژه و وجود ادبیات گفتاری و نوشتاری خاص که این روزه در شبکه های اجتماعی رواج دارند امکان استفاده مستقیم از این فایل ها وجود نداردو نیاز است برای استفاده در پروژه شخصی سازی شود ، در این مسیر ابتدا با بررسی تک تک توییت ها که در این پروژه آنها را یک سند در نظر می گیریم به نحوه نوشتار آن دقت می کنیم و فایل های پیش پردازش را بر طبق آن تغییر می دهیم یا از ابتدا آن ها را می نویسیم.

در قدم بعدی با توجه به اینکه پروژه عمیقا در زبان فارسی پژوهش انجام می دهد اکیدا نیاز است دیتاست را از هرگونه الفبای غیر فارسی پاکسازی کنیم که در این موضوع نیز با توجه به دیتاست نیاز است بررسی کنیم در توییت هایی که الفبای غیر فارسی در آن است با حذف تمامی کاراکتر های غیر فارسی مفهوم جمله چطور تغییر می کند؟ پس در این موضوع نیز نیاز است با بررسی دقیق تر اقدام به حذف الفبای غیر فارسی نمود.

در دیتاست اصلی با در سطر اول نام ویژگی های هر سطر آورده شده است، نیاز است در اولین اجرا و اولین اقدامی که صورت می گیرد آنهارا از دیتاست خارج کنیم، راهکار های دیگری که وجود دارد شامل نادیده گرفتن سطر اول است ، یا برنامه به نوعی بهینه شود که در صورت مواجه شدن با اطلاعاتی که در قالب خواسته شده نباشد انرا تشخیص دهد و آنرا در محاسبات لحاظ کند. قطعا برنامه ای که بتواند در همه شرایط بدون خطا اجرا شود و توانایی برخورد با انواع اطلاعات وارد نشده را داشته باشد از نظر عملکرد و پایداری در سطح بالاتری قرار می گیرد و در محیط های عملیاتی بار هزینه مالی و زمانی کمتری را بر دوش متقاضی برنامه می گذارد و همچنین برای برنامه نویس نیز این مسئله حائض اهمیت است که تا بالاترین حد امکان برنامه بهینه اجرا شود.

مورد دیگری که در این پروژه وجود دارد برخورد با اطلاعاتی هست که پس از اعمال فیلتر ها دیگر قابل استفاده نیست، تصور کنید با اعمال کلمات توقف و حذف حروف غیر فارسی به جمله ای برسیم که دارای تنها یک یا دو

کلمه باشد، مسلما اطلاعاتی که می توانیم از این متن دریافت کنیم کم و ناچیز خواهد بود و در ابعاد بزرگتر می تواند کارایی مدل مارا پایین بیاورد.

بدین جهت در متن برنامه تنها توییت هایی را مورد بررسی قرار می دهیم که دارای کلمه بیشتر از سه باشند و در صورت مواجهه با توییت هایی که که کمتر از مقدار گفته شده باشد آنرا نادیده خواهیم گرفت.

از چالش هایی که در این مسیر می توان به آن اشاره کرد بار محاسباتی حساب کردن مقدار tf-idf است، از جمله اقداماتی که در جهت کنترل کردن این مورد در برنامه لحاظ شده است ایجاد کردن لیست کلمات است، بدین صورت که از اولین خط خواندن دیتاست تمامی کلماتی که به آن برخورد می کنیم را به یک لیست اضافه می کنیم، این نکته قابل توجه است که تنها زمانی کلمه را به لیست اضافه می کنیم که آن کلمه در لیست ما وجود نداشته باشد و بدین ترتیب به لیستی دسترسی خواهیم داشت که حاوی تمام کلمات داخل دیتاست بدون در نظر گرفتن تکرار است، این امر باعث می شود که ما در طول برنامه و محاسبه مقادیر برای کل دیتاست، هیچ گاه اقدام به بررسی مقداری تکراری نکنیم، با توجه به بار محساباتی ای که گفته شد همین اقدام می تواند پیچیدگی زمانی پروژه را تا حد بسیار بالایی بکاهد و سرعت عمل کد ما را بالا ببرد.

با این حال پیچیدگی زمانی بالا کماکان در برنامه وجود خواهد داشت. اقدامی که در این پروژه انجام شد بررسی انواع روش های ممکن به جهت تسریع این فرایند ها بود که بهترین اقدام ممکن ذخیره مدل تولید شده در فایل می باشد.

بدین صورت که با اجرای برنامه و پس از انجام پیش پردازش ها برنامه دیکشنری مربوط به مقدار tf-idf کلمات به تفکیک شش احساس، آن را ذخیره می کنیم.

با توجه به اینکه در این مرحله از پروژه با یک دیکشنری مواجه هستیم باید روش صحیحی به جهت ذخیره آن انتخاب کنیم.

این روش انتخابی باید ویژگی های حداقلی را دارا باشد که می توان به موارد ذیل اشاره نمود.

نخست اینکه دلیلی که ما به ذخیره دیکشنری روی آوردیم کم کردن پیچیدگی زمانی این فرایند است، بنابراین با توجه به اینکه حجم داده های ما بسیار بالا می باشد باید روشی اتخاذ شود که پیچیدگی زمانی حداقلی را نیز دارا باشد. از طرف دیگر برنامه به خودی خود حجم بالایی از رم را اشغال می کند بنابراین نباید روش ذخیره سازی ما با فدا کردن پیچیدگی مکانی تنها پیچیدگی زمانی را بهبود ببخشد. روشی که ما در این پروژه از بهره جویی کردیم استفاده از توابع داخلی پایتون به جهت نوشتن و خواندن ساختمان داده های خود پایتون است که این کار را به نحو احسن انجام می دهند.

زمانی که ما در برنامه از توابع آماده خود پایتون استفاده می کنیم به این معنا است که این برنامه در بهترین حالت ممکن اجرا می شود و این موضوع می تواند اطمینان حاصل کند که بهنیه تری روش ممکن استفاده شده است.

۵-۲ پیشنهادات

اقداماتی که در جهت بهتر شدن این مدل که در این پروژه استفاده شده است می توان انجام داد به بهینه کردن الگوریتم محاسبه TF-IDF می توان اشاره کرد. این معیار امتیاز دهی اصلی ترین معیاری است که ما در این پروژه استفاده کرده ایم. مشخص است که هرچه این الگوریتم بهینه تر عمل کند جوابی که در انتها خواهیم داشت نیز بهینه خواهد بود، برای مثال می توانیم مقداری وزن جملات را نرمال کنیم که یکباره یک توییت با وزن بالا تاثیر کمتری در کلمات غیر مرتبط بگذارد.

با توجه به مورد گفته شده در بخش قبلی می توان در راستای بهبود عملکرد همین مدل انتخاب شده ینی استفاده از TF-IDF وزن دار اقدام به بالابردن کارایی برنامه کنیم، مواردی که می توان بطور مستقیم به آن اشاره کرد کلماتی است که به عنوان کلمه توقف از متن حذف می شوند.

با بهینه کردن هرچه بیشتر این کلمات می توانیم بازدهی بیشتری در برنامه داشته باشیم.

باید توجه داشت که مدل استفاده شده در این پروژه تا حد معقولی نتیجه صحیح را باز می گرداند و با تغییر جزئیات در پیش پردازش و نوع الگوریتم بهبود زیادی مشاهده نخواهد شد و مورد انتظار است زیر ده درصد بهبود حاصل شود.

به جهت رسیدن به کارایی بهتر نیاز است دیتاست اولیه ای که روی آن عملیات محاسبه انجام می دهیم به درستی وزن دهی شده باشد و در حالت پیش فرض و فرم اولیه اش نیز نیاز به پیش پردازش خاصی نداشته باشد، هرکجا که پیش پردازش انجام دهیم ممکن است مقداری از کارایی را از دست بدهیم.

استفاده از روش‌های ماشین لرنینگ برای تشخیص احساسات در متون

در عصر اطلاعات و فناوری، تحلیل احساسات در متون به ویژه در زمینه ادبیات فارسی، به یکی از چالش‌های اصلی پژوهشی تبدیل شده است. استفاده از روش‌های ماشین لرنینگ، به خصوص شبکه‌های عصبی و ماشین‌های بردار پشتیبان، در این حوزه امکانات گسترده‌تری برای تشخیص احساسات و اندازه‌گیری میزان هر احساس ارائه می‌دهد.

استفاده از شبکه‌های عصبی

شبکه‌های عصبی به دلیل توانایی بالا در یادگیری الگوهای پیچیده، برای تحلیل احساسات در متون ادبیات فارسی بسیار مناسب هستند. این شبکه‌ها می‌توانند اطلاعات احساساتی متن‌ها را به صورت خودکار و بدون نیاز به الگوهای دقیق، استخراج کنند. با آموزش این شبکه‌ها بر روی دیتاستی که شامل متون با برچسب‌های احساسات مختلف است، می‌توان به تشخیص و اندازه‌گیری احساسات در متون پرداخت.

استفاده از ماشین‌های بردار پشتیبان

ماشین‌های بردار پشتیبان نیز از قدرتمندترین روش‌های ماشین لرنینگ برای تحلیل احساسات متنی هستند. این روش با توجه به توانایی خود در دسته‌بندی داده‌ها، می‌تواند برای تفکیک احساسات مثبت، منفی، یا حتی احساسات میانه در متون ادبیات فارسی به کار گرفته شود. با تعیین یک مدل بهینه و استفاده از ویژگی‌های مناسب، می‌توان به تشخیص احساسات با دقت بالا دست یافت.

مقایسه با روش‌های بازیابی اطلاعات و TF-IDF

استفاده از شبکه‌های عصبی و ماشین‌های بردار پشتیبان برای تشخیص احساسات در مقایسه با روش‌های بازیابی اطلاعات و TF-IDF، اغلب نتایج بهتری ارائه می‌دهد. زیرا این روش‌ها قابلیت یادگیری و تفکیک بهتری را دارند و می‌توانند الگوهای پیچیده‌تری را در متون ادبیات فارسی تشخیص دهند. در حالی که روش‌های سنتی معمولاً بر اساس ویژگی‌های ساختاری یا تکراری متن عمل می‌کنند، ماشین لرنینگ با توانایی یادگیری و استفاده از ویژگی‌های پیچیده‌تر به دقت بیشتری دست می‌یابد.

استفاده از ماشین لرنینگ در تشخیص احساسات متن‌ها، نه تنها به افزایش دقت در تحلیل احساسات ادبیات فارسی منجر می‌شود، بلکه امکانات جدیدی برای تحقیقات آینده در زمینه متن کاوی و ادبیات فارسی ارائه می‌دهد.

مراجع

[1] N. Sabri, R. Akhavan, and B. Bahrak, “Emopars: A collection of 30k emotion-annotated persian social media texts,” in *Proceedings of the Student Research Workshop Associated with RANLP 2021*, 2021, pp. 167–173.

[2] S. S. Sadeghi, H. Khotanlou, and M. Rasekh Mahand, “Automatic Persian text emotion detection using cognitive linguistic and deep learning,” *J. AI Data Min.*, vol. 9, no. 2, pp. 169–179, 2021.

[3] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. USA: Cambridge University Press, 2008.

[4] H. Mirzaee, J. Peymanfard, H. H. Moshtaghin, and H. Zeinali, “ArmanEmo: A Persian Dataset for Text-based Emotion Detection,” *arXiv Prepr. arXiv2207.11808*, 2022.