بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی هوش مصنوعی

دسته‌بندی پاسخ‌های یک مکالمه اتوماتیک به مجموعه محدود با استفاده از تکنیک‌های هوش مصنوعی

نگارنده: مولود آیت

استاد راهنما

دکتر مرتضی زاهدی

بهمن ۹۶

نه می‌توانم موهایشان را که در راه عزت من سفید کرده‌اند را سیاه کنم و نه برای دست‌های پینه بسته‌شان که ثمره تلاش برای موفقیت من بود، مرهمی دارم.

پس پروردگارا توفیقم ده که هر لحظه شکرگزارشان باشم و ثانیه‌های عمرم را در عصای دست بودنشان

بگذرانم.

خدای را که سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت های او ندانند و کوشندگان، حق او را گزاردن نتوانند. و سلام و درود بر محمّد و خاندان پاک او، طاهران معصوم، هم آنان که وجودمان وامدار وجودشان است؛ و نفرین پیوسته بر دشمنان ایشان تا روز رستاخیز...

از پدر و مادر عزیزم... این دو معلم بزرگوارم... که همواره بر کوتاهی و درشتی من، قلم عفو کشیده و کریمانه از کنار غفلت هایم گذشته اند و در تمام عرصه های زندگی یار و یاوری بی چشم داشت برای من بوده اند؛ بسیار سپاسگزارم

بدون شک جایگاه و منزلت معلم، اجلّ از آن است که در مقام قدردانی از زحمات بی‌شائبه ی او، با زبان قاصر و دست ناتوان، چیزی بنگاریم.

از استاد با کمالات و شایسته؛ **جناب آقای دکتر مرتضی زاهدی** که در کمال سعه صدر، از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این رساله را بر عهده گرفتند؛

و از استاد فرزانه و دلسوز؛ **جناب آقای مهندس محمد مهدی حسینی** که بدون مساعدت ایشان، این رساله به نتیجه مطلوب نمی‌رسید؛ کمال تشکر و قدردانی را دارم.

همچنین لازم می‌دانم از دوست و هم‌کلاسی‌های عزیزم **سرکار خانم فاطمه زکی-زاده، غزاله مقدم‌نژاد** و **الهه علی‌پور** و دانشجویان آزمایشگاه وب‌کاوی و شناسایی الگو برای مساعدت‌ها و همراهی‌هایشان قدردانی نمایم.

باشد که این خردترین، بخشی از زحمات آنان را سپاس گوید.

تعهد نامه

اینجانب **مولود آیت** دانشجوی دوره کارشناسی ارشد رشته هوش مصنوعی دانشکده کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **دسته‌بندی پاسخ‌های یک مکالمه اتوماتیک به مجموعه محدود با استفاده از تکنیک‌های هوش مصنوعی** تحت راهنمائی **آقای دکتر مرتضی زاهدی** متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام‌شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورداستفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود است و مقالات مستخرج با نام «دانشگاه صنعتی شاهرود» و یا «Shahrood University of Technology» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافتهای آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

امضای دانشجو **تاریخ**

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود است. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

موتورهای جستجو کنونی قادر به پاسخ‌گویی مستقیم نیاز کاربران نیستند و معمولا لیستی از منابع مرتبط را برای آنها نمایش می‌دهند. فراهم کردن اطلاعات از طریق تعامل، ایده اصلی سیستم پرسش و پاسخ است. بیشتر تحقیق‌هایی که بر روی این سیستم‌ها انجام می‌شود، بر روی پردازش پرسش‌ها تمرکز دارند. پردازش پاسخ، تنها محدود به انتخاب پاسخ مناسب برای پرسش قبلی می‌شود. در این پژوهش دسته‌بندی قطبی بر روی پاسخ‌های یک سیستم پرسش و پاسخ تعاملی اعمال می‌شود. دسته‌بندی قطبی یکی از شاخه‌های مشهور در پردازش متن است که متن‌های ورودی را با توجه به محتوای درونی آن‌ها، به سه سطح مثبت، منفی و خنثی دسته‌بندی می‌کند.

انتخاب ویژگی‌های مناسب برای دسته‌بندی، یک امر بسیار مهم است. در این تحقیق دو ویژگی جدید با عنوان ویژگی احتمالی و اطلاعات متقابل نقطه به نقطه، معرفی می‌شوند. این ویژگی‌ها با توجه به گرایش بیشتر آن‌ها به یکی از سه دسته مثبت، منفی و خنثی، عبارت‌ها را مقداردهی می‌کنند. علاوه بر این دو ویژگی، سه ویژگی مشهور و متداول در دسته‌بندی قطبی نیز آزمایش می‌شوند. این آزمایش‌ها بر روی unigram ها و bigram ها انجام می‌شود. برای کاهش ابعاد نیز روش‌های بهره اطلاعات، نرخ بهره و مربع کای، اعمال می‌شوند.

تاکنون دسته‌بندی قطبی بر روی متن‌های تعاملی انجام نشده است. در یک مکالمه گفتگو پیشین می‌تواند در دسته‌بندی قطبی پاسخ، تاثیر گذارد. در این مطالعه، تاثیر گفته‌های پیشین به دو صورت وابستگی زمانی و وابستگی ساختاری بررسی خواهد شد. وابستگی زمانی محتوای گفته‌های پیشین را در دسته‌بندی تاثیر می‌دهد؛ درحالی که وابستگی ساختاری بر روی نوع آن‌ها تمرکز می‌کند. بهترین دقت با ویژگی احتمالی مبتنی بر unigram، برابر ۷۷/۴۵٪ است. با اعمال کاهش ویژگی، دقت ۷۵/۶۶٪ شد و با اعمال وابستگی ساختاری، مقدار آن به ۸۲/۶۲٪ رسید.

کلمات کلیدی: دسته‌بندی قطبی پاسخ، سیستم پرسش و پاسخ تعاملی، n-گرم، تاثیر گفتگوی پیشین

لیست مقالات مستخرج از پایان‌نامه

[۱] آیت، م.، حسینی، م م.، زاهدی، م.،(۱۳۹۶)، دسته‌بندی قطبی پاسخ‌ها در یک سیستم پرسش و پاسخ تعاملی، سومین کنفرانس پردازش سیگنال و سیستم‌های هوشمند، ص ۷۰، شاهرود

فهرست مطالب

عناوین **شماره صفحه**

فصل اول: مقدمه و مفاهیم پایه

۱-۱- مقدمه ۲

۱-۲- شکل‌های مختلف پاسخ‌گویی کاربر و سیستم ۲

۱-۳- چالش‌های موجود در تحلیل پاسخ‌های کاربران ۴

۱-۴- هدف پایان‌نامه ۵

۱-۵-ساختار پایان‌نامه ۶

۱-۶- جمع‌بندی ۶

فصل دوم: روش‌های متداول در دسته‌بندی قطبی

۲-۱- مقدمه ۸

۲-۲- مروری بر تحقیقات پیشین ۸

۲-۳- سطوح تحلیل در دسته‌بندی قطبی ۱۱

۲-۴- تعیین ویژگی‌های مناسب ۱۲

۲-۴-۱- حضور و تکرار عبارت‌ها ۱۳

۲-۴-۲-فرکانس معکوس سند ۱۴

۲-۴-۳- وزن‌دهی Tf-Idf ۱۴

۲-۴-۴- ویژگی‌های n-gram ۱۴

۲-۴-۵- منفی کننده‌ها ۱۵

۲-۵- شناسایی و برچسب زدن اجزای واژگان زبان ۱۵

۲-۵-۱- جداسازی صفت‌ها ۱۶
۲-۵-۲-ترکیب صفت‌ها و قیدها ۱۶
۲-۶- جهت معنایی عبارات ۱۶
۲-۶-۱- اطلاعات متقابل نقطه به نقطه ۱۷
۲-۷- روش‌های استخراج ویژگی ۱۸
۲-۷-۱- اطلاعات متقابل ۱۹
۲-۷-۳- بهره اطلاعات ۲۰
۲-۷-۴- نرخ بهره ۲۰
۲-۷-۵- مربع کای ۲۱
۲-۸- روش‌های مشهور در دسته‌بندی قطبی ۲۱
۲-۸-۱- روش‌های مبتنی بر یادگیری ماشین ۲۲
۲-۸-۲-تطابق الگو ۲۳
۲-۹- جمع‌بندی ۲۴

فصل سوم: سیستم‌های پرسش و پاسخ تعاملی

۳-۱- مقدمه ۳۰
۳-۲- خصوصیات گفتگوی انسانی ۳۰
۳-۲-۱- نوبت گرفتن ۳۱
۳-۲-۲-زمینه ۳۱
۳-۲-۳- ساختار مکالمه ۳۲
۳-۲-۴-رهبری مکالمه ۳۲

۳-۳- انواع مکالمه تحت وب ۳۳
۳-۴- سیستم‌های گفتگو ۳۴
۳-۵- معماری سیستم‌های گفتگو ۳۴
۳-۵-۱- پردازش ورودی ۳۵
۳-۵-۲- فهم زبان طبیعی ۳۶
۳-۵-۳- مدیر گفتگو ۳۶
۳-۵-۴- مولفه دامنه خاص ۳۷
۳-۵-۵- مولفه تولید پاسخ ۳۷
۳-۵-۶- انتشار خروجی ۳۷
۳-۶- طبقه‌بندی سیستم‌های گفتگو ۳۸
۳-۶-۱- سیستم‌های گفتگو مبتنی بر حالت محدود ۳۸
۳-۶-۲- سیستم مبتنی بر قالب ۳۹
۳-۶-۳- سیستم عامل مبتنی بر عامل ۴۰
۳-۷- سیستم‌های پرسش و پاسخ ۴۰
۳-۸- معماری سیستم‌های پرسش و پاسخ ۴۱
۳-۸-۱- پردازنده سوال ۴۱
۳-۸-۲- پردازنده سند ۴۲
۳-۸-۳- استخراج پاسخ ۴۲
۳-۹- سیستم پرسش و پاسخ تعاملی ۴۳
۳-۹-۱-انواع سیستم‌های پرسش و پاسخ تعاملی ۴۴

۳-۹-۲-اعلان مکالمه ۴۴

۳-۹-۳- ارتباط بین نوبت‌ها در مکالمه ۴۵

۳-۹-۴-ارزیابی ۴۵

۳-۱۰- جمع‌بندی ۴۷

فصل چهارم: استفاده از گفتگوی پیشین در دسته‌بندی قطبی

۴-۱- مقدمه ۵۰

۴-۳- پیش‌پردازش ۵۱

۴-۴- ویژگیهای انتخابی ۵۲

۴-۴-۱- احتمال رخداد یک عبارت در یک دسته (ویژگی احتمالی) ۵۳

۴-۴-۲- اطلاعات متقابل نقطه به نقطه ۵۴

۴-۵- روش‌های دسته‌بندی ۵۴

۴-۵-۱- بیزین ساده ۵۴

۴-۵-۲-ماشین بردار پشتیبان ۵۵

۴-۵-۳- درخت تصمیم ۵۶

۴-۵-۴- الگوریتم Bagging (درخت تصمیم) ۵۶

۴-۶- روش‌های کاهش ویژگی ۵۷

۴-۶-۱ بهره اطلاعات ۵۷

۴-۶-۲- نرخ بهره ۵۸

۴-۶-۳- مربع کای ۵۸

۴-۷- بررسی سابقه گفتگو ۵۸

۴-۷-۱- وابستگی زمانی ۵۸
۴-۷-۲- وابستگی ساختاری ۵۹
۴-۸- روش‌های ارزیابی ۵۹
۴-۹- جمع‌بندی ۶۰

فصل پنجم: ارزیابی روش‌های پیشنهادی و بررسی نتایج

۵-۱- مقدمه ۶۲
۵-۲- پایگاه داده ۶۲
۵-۳- بررسی روش‌های پیش‌پردازش ۶۳
۵-۳-۱- پیش‌پردازش اولیه ۶۳
۵-۳-۲- ترکیب پیش‌پردازش اولیه با ریشه‌یابی ۶۶
۵-۳-۳-بررسی پیش‌پردازش با حذف کلمات اضافی ۶۸
۵-۴- بررسی روش‌های کاهش ویژگی ۷۰
۵-۴-۱-بررسی روش‌های کاهش ویژگی بر روی unigram ۷۱
۵-۴-۲-بررسی روش‌های کاهش ویژگی بر روی bigram ۷۲
۵-۴-۳-بررسی روش‌های کاهش ویژگی بر روی ترکیب unigram و bigram ۷۴
۵-۵- بررسی تاثیر گفته‌های پیشین در دسته‌بندی پاسخ‌ها ۷۶
۵-۵-۱- وابستگی زمانی ۷۷
۵-۵-۲- وابستگی ساختاری ۷۹
۵-۶-جمع‌بندی ۸۰

فصل ششم: جمع‌بندی و پیشنهادهایی برای ادامه کار

۶-۱- جمع‌بندی ۸۴
۶-۲- پیشنهادهایی برای ادامه کار ۸۵
مراجع ۸۶

فهرست اشکال

عناوین **شماره صفحه**

شکل (۱-۱): اشکال مختلف پاسخ‌گویی ۳

شکل (۲-۱): روش‌های دسته‌بندی قطبی ۲۱

شکل (۳-۱): مولفه‌های اصلی یک سیستم گفتگو ۳۵

شکل (۳-۲): مثالی از سیستم گفتگو مبتنی بر حالت محدود ۳۹

شکل (۳-۳): مثالی از سیستم گفتگو مبتنی بر قالب ۳۹

شکل (۳-۴): مولفه‌های اصلی سیستم پرسش و پاسخ ۴۱

شکل (۴-۱): فرایند دسته‌بندی پاسخ‌ها ۵۰

شکل (۴-۱): نحوه تاثیرگذاری نوبت‌های پیشین در پاسخ فعلی ۵۹

شکل(۵-۱): انتخاب ضریب a مناسب توسط دقت ماشین بردار پشتیبان (چندجمله‌ای و خطی) و ویژگی احتمالی ۷۷

شکل(۵-۲): انتخاب ضریب a مناسب توسط دقت ماشین بردار پشتیبان (چندجمله‌ای و خطی) و PMI ۷۸

شکل (۵-۳): مقایسه دو تابع هسته ماشین بردار پشتیبان ۸۰

فهرست جداول

عناوین **شماره صفحه**

جدول (۲-۱): مدل‌های مختلف فرکانس عبارت[۲۲] ۱۳

جدول (۲-۲): مدل‌های مختلف فرکانس معکوس سند [۲۲] ۱۴

جدول (۲-۳): برچسب‌های واژگان در اجزای واژگان زبان ۱۵

جدول (۲-۴): مقایسه روش‌های دسته‌بندی قطبی در منابع مختلف ۲۶

جدول (۵-۱): پیش‌پردازش اولیه و unigram ۶۴

جدول (۵-۲): پیش‌پردازش اولیه و bigram ۶۵

جدول (۵-۳): ترکیب پیش‌پردازش اولیه با ریشه‌یابی و unigram ۶۶

جدول (۵-۴): ترکیب پیش‌پردازش اولیه با ریشه‌یابی و bigram ۶۷

جدول (۵-۵): ترکیب پیش‌پردازش اولیه با حذف کلمات اضافی و unigram ۶۸

جدول (۵-۶): ترکیب پیش‌پردازش اولیه با حذف کلمات اضافی و bigram ۶۹

جدول (۵-۷): تأثیر روش‌های کاهش ویژگی بر روی unigram و ویژگی احتمالی ۷۱

جدول (۵-۸): تأثیر روش‌های کاهش ویژگی بر روی unigram و PMI ۷۱

جدول (۵-۹): تأثیر روش‌های کاهش ویژگی بر روی bigram و ویژگی احتمالی ۷۳

جدول (۵-۱۰): تاثیر روش‌های کاهش ویژگی بر روی bigram و PMI ۷۳

جدول (۵-۱۱): ترکیب unigram و bigram کاهش یافته و ویژگی احتمالی ۷۵

جدول (۵-۱۲): ترکیب ویژگی n-gram های کاهش یافته و PMI ۷۵

جدول(۵-۱۳): درصد دقت دسته‌بندی‌های مختلف با تاثیرگذاری وابستگی ساختاری ۷۹

فصل اول:

مقدمه و مفاهیم پایه

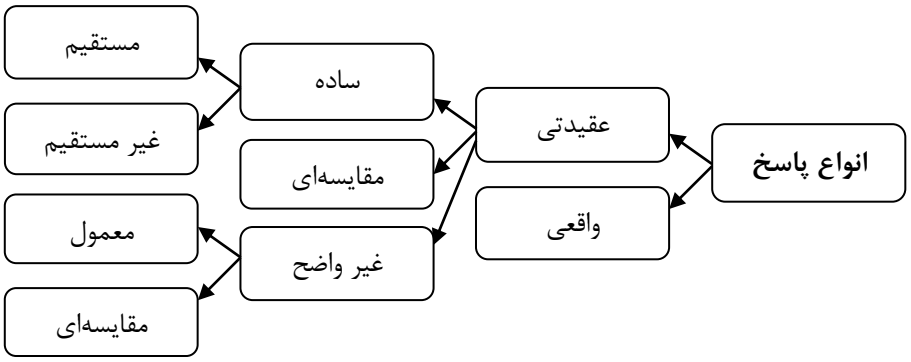
۱-۱- مقدمه

محیط‌های بحث مجازی به کاربران این امکان را می‌دهند که آزادانه افکار خود را بیان کنند. محتوای متن تولید شده توسط کاربر اهمیت حیاتی برای سازمان‌ها و دیگر کاربران دارد. از طرفی مردم می-توانند نظرات خود را در یک محیط بدون تبعیض و محدودیت به اشتراک بگذارند، خود را ابراز کنند، بر دیگران تاثیر گذارند، در جوامع مجازی عضو شوند و فضای مجازی خود را داشته باشند. از طرفی دیگر، سازمان‌ها می‌توانند بینش‌های مفیدی را برای فرایند تصمیم‌گیری خود به دست آورند. افزایش اطلاعات بدون ساختار و نیمه ساختاریافته کاربران در سطح اینترنت، مشکلاتی را بوجود آورده است. پردازش و سازماندهی این اطلاعات، هدفی چالش برانگیز است. زیرا برخی از این اطلاعات، کوتاه هستند و در پاسخ به یک موجودیت بیان شدند. همچنین حالت‌های مختلفی برای پاسخ‌گویی وجود دارد و یک پاسخ، به یک شکل واحد بیان نمی‌شود. در این فصل انواع حالت‌های مختلف پاسخ‌گویی و چالش‌های موجود در دسته‌بندی آن‌ها، بررسی خواهند شد؛ هدف پایان‌نامه توضیح داده می‌شود و در پایان ساختار پایان‌نامه معرفی می‌شود.

۱-۲- شکل‌های مختلف پاسخ‌گویی کاربر و سیستم

بیشتر روش‌های کلاس‌بندی سنتی متن، تفاوت‌های بسیار کم را در دسته‌بندی تاثیر نمی‌دهند. این روش‌ها بر این اساس کار می‌کنند که دو متن شبیه به هم ممکن است دارای تفاوت‌های جزئی باشند. دسته‌بندی قطبی، چالش برانگیزتر از روش‌های دسته‌بندی سنتی مانند دسته‌بندی موضوعی است. زیرا افراد همواره پاسخ‌های خود را به صورت یک شکل خاص بیان نمی‌کنند. به عنوان مثال دو جمله «من وقت دارم با شما در مورد این موضوع صحبت کنم» و «من وقت ندارم با شما در مورد این موضوع صحبت کنم» تفاوت آشکاری با یکدیگر دارند. علاوه بر این، ویژگی‌های احساسی به صورت عینی و صریح بیان نمی‌شوند و معمولا در یک مجموعه بزرگ از متن با احساس‌های درونی پنهان می‌شوند. بنابراین در دسته‌بندی قطبی، نیاز به تجزیه و تحلیل عمیق‌تر و درک ویژگی‌های متنی

است[۱]. برای بیان یک پاسخ، شکل‌های مختلفی وجود دارد که شکل (۱-۱) بعضی از این اشکال را نشان داده است.



شکل (۱-۱): اشکال مختلف پاسخگویی

پاسخ‌های واقعی و عقیدتی: پاسخ‌های واقعی، حقایقی در مورد یک موضوع خاص بیان می‌کنند. به عنوان مثال «اوباما انتخابات را برنده شد». این حقایق با توجه به گفتگو قبلی می‌توانند مطلوب و یا نامطلوب باشند. پاسخ‌های عقیدتی بیشتر جنبه احساسی دارند. به عنوان مثال جمله «به عقیده من اوباما رئیس‌جمهورخوبی می‌شود». در این حالت، ممکن است عقیده بعضی افراد مخالف این پاسخ باشد[۲]. پاسخ‌های عقیدتی به دو دسته ساده (مستقیم و غیرمستقیم) و مقایسه‌ای تقسیم می‌شوند.

پاسخ ساده مستقیم: در این نوع پاسخ، گوینده احساس خود را درباره جنبه‌های یک موجودیت خاص بیان می‌کند. به عنوان مثال در جمله «پنجره‌ای در این اتاق نیست و این باعث می‌شود اتاق تاریک شود»، موجودیت در این پاسخ اتاق و جنبه، پنجره است.

پاسخ ساده غیر مستقیم: در این نوع پاسخ به صورت غیر مستقیم ویژگی‌های یک موجودیت بیان می‌شود. به عنوان مثال «این خانه نوساز است». تشخیص این نوع نظر تا حدودی مشکل است زیرا در بعضی از مواقع می‌توان آن را با پاسخ واقعی اشتباه گرفت. در این حالت با توجه به گفته‌های پیشین دسته این نوع پاسخ مشخص می‌شود.

پاسخ مقایسه‌ای: یک پاسخ مقایسه‌ای، تفاوت‌ها و شباهت‌های بین دو یا چند موجودیت را بیان می‌کند. به عبارت دیگر جنبه‌های مختلف دو موجودیت را با یکدیگر مقایسه می‌کند. به عنوان مثال «این خانه نسبت به قبلی، فضای بزرگتری دارد».

پاسخ غیر واضح: یک شرح بدیهی است که براساس واقعیت‌های مطلوب یا غیرمطلوب که به دو صورت معمول و مقایسه‌ای بیان می‌شود. به عنوان مثال «نمی‌توانم تا شب برای والیبال صبر کنم» و «خشونت در فیلم نجات سرباز رایان بیشتر از شکارچی گوزن است». در مورد مثال اول پیش‌بینی‌های مثبت می‌شود که به صورت واضح بیان نشده و در پاسخ دوم تشخیص قطب، حتی برای خود انسان‌ها نیز مشکل است. برای بعضی افراد فیلم‌هایی که حاوی خشونت بیشتری هستند یک مشخصه مثبت است زیرا موجب می‌شود فیلم به واقعیت نزدیک‌تر شود. این درحالی است که برای افرادی دیگر، این مشخصه منفی باشد.

۱-۳- چالش‌های موجود در تحلیل پاسخ‌های کاربران

علاوه بر اشکال مختلف پاسخ، چالش‌های زیاد دیگری در پردازش آن‌ها وجود دارد. این پاسخ‌ها ویژگی‌های خاصی دارند که آن‌ها را با دیگر متن‌ها متفاوت می‌سازد. به طور کلی این ویژگی‌ها عبارتند از:

کلمات احساسی: وجود این نوع کلمه‌ها در پاسخ، در بعضی موقعیت‌ها معنی پاسخ را مثبت و در بعضی دیگر مفهوم پاسخ را منفی می‌کنند. این کلمه‌ها به طور کلی نمایانگر احساس مربوط به یک جنبه خاص از یک موجودیت است.

دوقطبی بودن پاسخ‌ها: ممکن است در یک پاسخ، هر دو نوع جمله مثبت و منفی موجود باشد. در این حالت بایستی دسته پاسخ را باتوجه به گرایش بیشتر آن به یکی از قطب‌ها، انتخاب کرد. به عنوان مثال در پاسخ مکالمه زیر پاسخ هم مثبت است و هم منفی ولی گرایش پاسخ به سمت منفی بیشتر است.

- آیا از اتاق راضی هستین؟
- اتاق امکانات زیادی دارد اما کثیف است.

مربوط بودن به گفتگو قبلی: در اکثر مواقع مثبت یا منفی بودن یک پاسخ، به جملات گفته شده قبلی مربوط می‌شود. به عنوان مثال در مکالمه زیر:

- آیا درآمد دیگری دارید که به خرید خانه کمک کند؟
- ماه پیش هزار دلار از بیمه گرفتم

پاسخ این سوال مثبت است ولی در ظاهر و بدون در نظر گرفتن گفتگو قبلی، این پاسخ خنثی در نظر گرفته می‌شود.

مربوط بودن به حوادث و اتفاق‌ها در آینده: در بعضی موارد پاسخ کاربر در صورت وقوع حوادث بعدی مثبت یا منفی می‌شود. این پاسخ‌ها معمولا به صورت یک شرط بیان می‌شوند. به عنوان مثال «در صورتی که قیمت این خانه متناسب با بودجه ما باشد، این خانه را می‌خریم».

۱-۴- هدف پایان‌نامه

هدف اصلی این پژوهش، دسته‌بندی قطبی پاسخ‌های یک سیستم پرسش و پاسخ تعاملی است. سیستم پرسش و پاسخ تعاملی از طریق مکالمه اطلاعات مورد نیاز کاربر را فراهم می‌کند. پردازش پاسخ در اینگونه سیستم‌ها محدود به انتخاب پاسخ مناسب برای پرسش نوبت قبلی می‌شود. علاوه بر دسته‌بندی پرسش‌ها، دسته‌بندی پاسخ‌ها نیز می‌تواند اطلاعات مهمی را فراهم کند. این اطلاعات می‌تواند به انتخاب گفته‌های بعدی کمک کند. همچنین می‌تواند میزان رضایت کاربر و سیستم را از گفتگو پیشین، نمایش دهد. با دسته‌بندی قطبی پاسخ‌ها، می‌توان تفکر کاربران و سیستم را به صورت کیفی اندازه‌گیری کرد. تاکنون روش‌هایی که برای دسته‌بندی قطبی استفاده شده است، متن‌ها را بدون در نظر گرفتن سابقه آنها دسته‌بندی کرده‌اند. بسیاری از متن‌ها در پاسخ به متنی دیگر بیان شده‌اند. این ارزیابی کیفی به انتخاب جمله‌های بعدی کمک می‌کند و با انتخاب جمله‌های مناسب، کاربر سریع‌تر می‌تواند اطلاعات مورد نیاز خود را پیدا کند. می‌توان گفت که این پژوهش، اولین کاری

است که تاثیر سابقه متن را در دسته‌بندی قطبی تاثیر می‌دهد. همچنین نخستین تحقیقی است که به دسته‌بندی پاسخ‌های یک سیستم پرسش و پاسخ تعاملی می‌پردازد.

۱-۵-ساختار پایان‌نامه

این پایان‌نامه در شش فصل تدوین شده است. در فصل اول به توضیحاتی در مورد کلیات این پژوهش پرداخته شد؛ فصل دوم ویژگی‌ها و روش‌های مشهور در دسته‌بندی قطبی و بعضی از روش‌های مورد آزمایش، معرفی می‌شوند. ساختار مکالمه، نحوه تعامل و سیستم‌های که از مکالمه برای فراهم کردن اطلاعات مورد نیاز کاربر استفاده می‌کنند، در فصل سوم بررسی می‌شوند. فصل چهارم و پنجم به معرفی روش پیشنهادی و نتایج آزمایش‌ها پرداخته می‌شود. در نهایت در فصل ششم، برای بهبود مکالمه بین کاربر و سیستم روش‌هایی، پیشنهاد می‌شود.

۱-۶- جمع‌بندی

در این فصل مفاهیم پایه و هدف اصلی پایان‌نامه ذکر شد؛ حالت‌های مختلفی که برای پاسخگویی وجود دارند، بررسی شدند؛ همچنین چالش‌هایی که در دسته‌بندی آن‌ها وجود داشت نیز شرح داده شدند. بررسی گذشته تحقیقات مشابه با این زمینه، یکی از بخش‌های مهم این فصل بود که در بخش (۴-۱) به آن پرداخته شد. این بررسی، دید کلی در مورد روش‌های محققان در اختیار قرار می‌دهد که در فصل‌ها بعدی معرفی خواهند شد و در پایان ساختار پایان‌نامه معرفی شد.

فصل دوم

روش‌های متداول در دسته‌بندی قطبی

۲-۱- مقدمه

اینترنت در زندگی مردم یک انقلاب دیجیتالی ایجاد کرده است. امروزه، افراد قادر هستند احساسات و عقیده خود را بدون هیچ محدودیتی بیان کنند. مهمتر از همه، به اشتراک‌گذاری و گسترش افکار خود با افراد دیگر بدون هیچ گونه مانع جغرافیایی بپردازند. از این رو، این انقلاب منجر به افزایش وظایف تجزیه و تحلیل احساسات[۱] شده است. مدل‌های تجزیه و تحلیل احساسات، نیاز به در نظر گرفتن یک محیط مجازی (جایی که کاربران در تعامل با یکدیگر) دارند و باید تفکر مردم و آنچه آنها در نظر دارند را به تصویر بکشند. دسته‌بندی قطبی یکی از شاخه‌های مشهور در تجزیه و تحلیل احساسات است که متن‌های ورودی را با توجه به محتوای درونی آن‌ها، به سه سطح مثبت، منفی و خنثی دسته‌بندی می‌کند.

در این فصل روش‌های مشهور در دسته‌بندی قطبی بررسی می‌شوند. در ابتدای کار سابقه تحقیق به صورت جزئی معرفی می‌شود؛ سپس سطح‌های مختلف دسته‌بندی قطبی گفته می‌شود. در مرحله‌ی بعد ویژگی‌های موثر و مشهور در دسته‌بندی قطبی توضیح داده می‌شوند. سپس روش‌های کاهش ویژگی پرکاربرد در این حوزه، شرح داده می‌شوند و در پایان روش‌ها و تکنیک‌های موثر در این زمینه بررسی خواهد شد.

۲-۲- مروری بر تحقیقات پیشین

پیش از شروع وب ۲٫۰ کاربران اینترنتی تنها نقش مصرف‌کننده اینترنت را داشتند و نمی‌توانستند دیدگاه‌های خود را آزادانه بیان کنند. وب ۲٫۰ نقطه آغاز تعامل میان کاربر و اینترنت است. وبلاگ‌ها، انجمن‌ها، چت روم‌ها، شبکه‌های تجاری و اجتماعی نمونه منابعی هستند که براساس تعامل ایجاد شدند. کاربران اینترنتی بوسیله این منابع می‌توانند به آسانی با یکدیگر تعامل کنند و علایق و منابع خود را با دیگران به اشتراک بگذارند[۳]. حتی برخی از سایت‌هایی که از فرمت قدیمی ایستا استفاده می‌کردند مانند روزنامه‌ها، نیز حالت اجتماعی پیدا کرده و کاربران می‌توانند نظر خود را بعد از هر خبر

[۱] Sentiment Analysis

به اشتراک بگذارند[۴]. اوایل سال ۲۰۰۰ شروع تحقیقات بر روی تعاملات کاربران در شبکه‌های اجتماعی بود که با عنوان آنالیز احساسی شناخته می‌شود[۵]. آنالیز احساسی یکی از شاخه‌های مدیریت اطلاعات است که از روش‌های پیشرفته پردازش متن، یادگیری ماشین، بازیابی اطلاعات و پردازش زبان طبیعی برای پردازش محتوای متن تولید شده توسط کاربر استفاده می‌کند[۱]. هدف آنالیز احساسی، استخراج اطلاعات اساسی از متون مختلف است. براساس کاربرد، نام‌های متفاوتی برای آن وجود دارد مانند پردازش عقیده، استخراج عقیده، پردازش احساس، تحلیل موضوع، تحلیل تاثیر، تحلیل هیجان و پردازش بازدید. واضح است که این کار برای همه زبان‌ها اهمیت دارد؛ با این حال، شکاف قابل ملاحظه‌ای بین روش‌های پیشرفته انگلیسی و زبان های دیگر وجود دارد. همانطور که انتظار می‌رود، برخی از محققان تصمیم می‌گیرند تا مستقیما دیدگاه‌ها را به زبان انگلیسی ترجمه کنند و سپس از یک ابزار آنالیز احساسی انگلیسی قوی استفاده کنند [۶, ۷]. برخی دیگر یک طبقه بند احساسی صحیح برای زبان مورد نظر ارائه می‌دهند [۸].

معمولاً آنالیز احساس به صورت نادرست به دسته‌بندی قطبی[۱] تعبیر می‌شود. دسته‌بندی قطبی یکی از وظایف آنالیز احساسی است که به طور کلی احساس مثبت، منفی و خنثی (که به قطب یا بار معروف هستند) را از متون مختلف استخراج می‌کند[۵]. برای دسته‌بندی قطبی سطح‌های مختلفی وجود دارد. بعضی از تحقیق‌ها تنها بر روی دو سطح مثبت و منفی تمرکز کردند[۹]؛ تعدادی بر روی سه سطح مثبت، منفی و خنثی[۱۰] و عده‌ای دیگر بر روی سطح‌های بالاتر مانند استفاده از ستاره [۱۱] تحقیق کردند.

یک مسئله مهم، دستیابی کاربران به اطلاعات مورد نیاز خود است که مدل‌های کلاسیک بازیابی اطلاعات قادر به استخراج آن‌ها نیستند. بازیابی اطلاعات به فرایند نمایش، ذخیره، سازماندهی و استخراج اطلاعات ذخیره شده در مجموعه داده‌های عظیم با توجه به نیاز کاربر گفته می‌شود[۱۲]. مدل‌های کلاسیک بازیابی اطلاعات، بعد اجتماعی را در سطح اینترنت در نظر نمی‌گیرند و الگوریتم-

[۱] Polarity Classification

های رتبه‌بندی آنها، تنها براساس شباهت بین پرسش مطرح شده و اسناد موجود، تعداد ارجاع به یک سند خاص و غیره متمرکز است. مشکل اساسی مدل‌های کلاسیک بازیابی اطلاعات در این است که کاربران بایستی به جستجوی اطلاعات مورد نیاز خود از بین سندهای رتبه‌بندی شده بپردازند و به طور مستقیم نمی‌توانند آن‌ها را بدست آورند. بدون تردید استفاده از اطلاعات اجتماعی در بازیابی اطلاعات مزیت‌هایی دارد. به عنوان مثال بازخورد اطلاعات می‌تواند مستقیما توسط کاربر بیان شود و کاربر می‌تواند میزان رضایت خود را با استفاده از عبارت‌هایی که استفاده می‌کند، نشان دهد[۱۳]. در بعضی موارد درخواست کاربر مبهم است در این حالت سیستم نیز می‌تواند با پرسش سوال‌هایی، ابهام-های موجود را برطرف کند و همچنین از کاربر بخواهد با ارائه توضیحات بیشتر، درخواست خود را تصحیح کند.

در بازیابی اطلاعات، پرسش و پاسخ به معنای پاسخگویی خودکار به یک سوال با استفاده از زبان طبیعی است [۱۴]. پرسش و پاسخ در چند سال اخیر توجه زیادی را از جامعه پژوهش محاسباتی زبان‌شناسی دریافت کرده است. به طور کلی سیستم‌های پرسش و پاسخ در سال ۱۹۶۰ توسعه یافتند. یکی از سیستم‌های اولیه که با نام Baseball شناخته می‌شود، سوال‌های مربوط به بازی بیس‌بال را براساس معانی آن‌ها تجزیه می‌کرد و پاسخ آن‌ها را در یک پایگاه‌داده ساختاریافته جستجو می‌کرد. این سیستم‌ها که دامنه‌ی محدودی داشتند تا سال ۱۹۸۰ ادامه پیدا کردند. در دهه ۹۰ گامی مهم در پردازش زبان طبیعی به وجود آمد و سیستم‌های پرسش و پاسخ با دامنه‌ی باز توسعه یافتند. این سیستم‌ها بر پردازش محدود زبانی و بازیابی اطلاعات تکیه می‌کردند و از داده‌های بدون ساختار متنی استفاده می‌شد. به عنوان مثال Murax یک سیستم پرسش و پاسخ متن باز است که روش‌های قدرتمند زبانی و یک سیستم بازیابی اطلاعات را ترکیب کرده است. این سیستم پرسش و پاسخ، پاسخ‌های مربوط به یک دانش‌نامه آنلاین را پیدا می‌کند[۱۵].

سیستم‌های گفتگو[۱] و سیستم‌های پرسش و پاسخ تعاملی[۲]، اطلاعات مورد نیاز کاربر را از طریق تعامل فراهم می‌کنند. نسخه اولیه سیستم‌های گفتگو، ربات‌های چت[۳] بودند که هدف آنها نسبتا ساده بود. نقش اولیه آن‌ها مشابه آزمون تورینگ بود. به بیانی دیگر، تلاش می‌کردند کاربر را با تفکر اینکه در حال ارتباط با انسان هستند، فریب دهند [۱۶]. به مرور علاقه به سیستم‌های گفتگو از تحقیقات علمی خالص به کاربرد تجاری تغییر پیدا کرد. سیستم‌های گفتگویی که توسط برخی از محققین [۱۷] به عنوان ماموران مکالمه[۴] نامیده می‌شوند، در زمینه‌هایی مانند خدمات مشتری، سرویس‌های خدماتی، هدایت وب سایت، هدایت فروش و پشتیبانی فنی، استفاده می‌شوند. علت به‌کارگیری آن‌ها ارائه یک روش طبیعی برای تعامل با کاربر است. به عبارت دیگر کاربران معمولا نیاز به آموزش ویژه ای ندارند. زیرا استفاده از آن‌ها آسان و شهودی هستند. ELIZA و PARRY نمونه سیستم‌های گفتگو اولیه هستند [۱۸]. پروژه QALL-ME نمونه اولیه یک سیستم پرسش و پاسخ تعاملی است که در حوزه گردشگری توسعه یافته است [۱۹]. برای رسیدن به این هدف، یک هستی‌شناسی برای گردشگری ایجاد شده و در روند پاسخگویی مورد استفاده قرار می‌گیرد. هر زمان که یک سوال با توجه به دامنه هستی‌شناسی نتواند تجزیه تحلیل یا پاسخ‌دهی شود، سیستم نتیجه‌گیری می‌کند که درخواست خارج از دامنه است.

۲-۳- سطوح تحلیل در دسته‌بندی قطبی

دسته‌بندی قطبی (مثبت، منفی و خنثی) پاسخ‌ها، یکی از وظایف آنالیز احساسی است. دسته‌بندی قطبی، متن‌ها را در سه سطح مختلف دسته‌بندی می‌کند.

سطح سند: در این سطح دسته‌بندی قطبی، متن به صورت کامل به یکی از سه دسته مثبت، منفی و خنثی دسته‌بندی می‌شود. یک مکالمه از چند نوبت تشکیل شده است. در هر نوبت یکی از افراد،

[۱] Dialogue System
[۲] Interactive Question Answering (IQA)
[۳] Chatterbots - Chatbots
[۴] Conversational Agents

عقیده و پاسخ خود را نسبت به گفتگوهای قبلی بیان می‌کند. یک عقیده و یا پاسخ، ممکن است از جمله‌های مختلفی تشکیل شده باشد. این جمله‌ها می توانند حاوی قطب‌های متفاوت و یا حاوی سوال باشد[۲]. در این سطح، پاسخ کاربر به صورت کلی نسبت به گفتگو پیشین و گرایش بیشتر آن به سه دسته مثبت، منفی و خنثی، کلاس‌بندی می‌شود. این سطح نام‌های دیگری مانند سطح پیام و پاسخ نیز دارد[۲, ۲۰]. در این پایان‌نامه دسته‌بندی قطبی پاسخ‌ها، در سطح سند انجام می‌شود.

سطح جمله: همانطور که قبلا ذکر شد پاسخ کاربر نسبت به یک موضوع خاص ممکن است از چندین جمله تشکیل شده باشد. دسته‌بندی قطبی در این سطح، هر یک از جمله‌ها را به صورت جداگانه کلاس‌بندی می‌کند. به بیانی دیگر این سطح جزئی‌تر از سطح پاسخ عمل می‌کند.

سطح موجودیت: این سطح جزئی‌تر از دو سطح دیگر کلاس‌بندی را انجام می‌دهد. در این سطح، دسته‌بندی قطبی، موجودیت مهمی که در مورد آن اظهارنظر شده را تشخیص می‌دهد و بر اساس جنبه‌های بیان شده، آن را کلاس‌بندی می‌کند. این روش بیشتر مورد استفاده خریداران و تولیدکنندگان اینترنتی است که علاقه‌مند هستند اطلاعات کاملی در مورد یک محصول بدست آورند[۲۰]. به عنوان مثال «لپ‌تاپ لنوو بسیار قدرتمند است اما سازندگان بایستی کیفیت تصویر و طول عمر باتری را بهبود دهند». در این پاسخ، موجودیت لپ‌تاپ لنوو(مثبت) و جنبه‌ها کیفیت تصویر(منفی) و طول عمر باتری (منفی) هستند.

۲-۴- تعیین ویژگی‌های مناسب

مهم‌ترین گام در دسته‌بندی قطبی، انتخاب ویژگی‌های مناسب جهت کلاس‌بندی است. عوامل بسیاری در انتخاب ویژگی‌ها تاثیرگذارند. طبق مرجع [۲۱]، یک ویژگی مناسب در یک محیط دسته-بندی بایستی حاوی خصوصیت‌های زیر باشد:

- یک ویژگی مناسب باید در محدوده مسئله، مهم و بامفهوم باشد.

- یک ویژگی باید نماینده کلاس مربوطه باشد. به عبارت دیگر، با استفاده از آن بتوان کلاس‌های مختلف را متمایز کرد.
- یک ویژگی باید نسبت به مشخصه‌های دیگر مستقل باشد.

در ادامه بعضی از عمومی‌ترین مشخصه‌های استفاده شده در دسته‌بندی قطبی بررسی می‌شوند.

۲-۴-۱- حضور و تکرار عبارت‌ها

تکرار عبارت، همواره به عنوان یک مشخصه مهم در بازیابی اطلاعات[۱] و دسته‌بندی متون درنظر گرفته می‌شود. علاوه بر تکرار، حضور عبارت نیز می‌تواند در دسته‌بندی متون کوتاه نقشی اساسی ایفا کند. علت این امر وجود تکرار کم عبارت‌ها در متون کوتاه است. حضور عبارت، یک بردار ویژگی دودویی است که در صورت مشاهده عبارت، مقدار یک و در غیر اینصورت مقدار صفر خواهد داشت. مرجع[۲۲] مدل‌های مختلف حضور و تکرار عبارت را جمع‌آوری کرده و در قالب جدول (۲-۱) نمایش داده است.

جدول (۲-۱): مدل‌های مختلف فرکانس عبارت[۲۲]

عنوان	فرمول
تکرار عبارت کلاسیک	tf
تکرار عبارت لگاریتمی	$1+\log(tf)$
تکرار عبارت تکمیلی	$0.5+\frac{0.5\times tf}{\max_t(tf)}$
تکرار عبارت دودویی (حضور عبارت)	$\begin{cases}1, & tf \succ 0\\ 0, & otherwise\end{cases}$
تکرار عبارت میانگین لگاریتمی	$\frac{1+\log(tf)}{1+\log(avg_{dl})}$

در تکرار عبارت تکمیلی، $\max_t(tf)$ ماکزیمم تکرار هر عبارت در متن‌ها را نشان می‌دهد و avg_{dl} میانگین تعداد عبارت‌ها در همه متن‌ها است.

[۱] Information Retrieval

۲-۴-۲-فرکانس معکوس سند

این ویژگی برای عبارت‌هایی که تکرار کمی را در متن دارند امتیاز بالا و برای عبارت‌هایی که تقریبا در هر متنی تکرار شده‌اند، امتیاز کمی را محاسبه می‌کند. حالت کلاسیک آن از تقسیم تعداد کل متن‌ها بر متن‌هایی که حاوی آن عبارت هستند، بدست می‌آید. به طور کلی این ویژگی برای دسته-بندی متون در بازیابی اطلاعات مبهم است. زیرا اطلاعاتی کلی در مورد توزیع یک عبارت بدون درنظر گرفتن نوع کلاس آن را فراهم می‌کند[۲۲]. برای محاسبه فرکانس معکوس سند روش‌های مختلفی وجود دارد. مرجع [۲۲] این روش‌ها را دسته‌بندی کرده و به صورت جدول شماره(۲-۲) نمایش داده است.

جدول (۲-۲): مدل‌های مختلف فرکانس معکوس سند [۲۲]

عنوان	فرکانس معکوس سند
حالت کلاسیک	$\log\frac{N}{df}$
حالت احتمالی	$\log\frac{N-df}{df}$

N، در جدول (۲-۲) تعداد کل متن‌ها را نشان می‌دهد؛ df تعداد متنی است که حاوی عبارت مورد نظر است.

۲-۴-۳- وزن‌دهی Tf-Idf

هنگامی که دو ویژگی تکرار عبارت و فرکانس معکوس سند در یکدیگر ضرب می‌شوند، مجموعه ویژگی جدیدی بوجود می‌آید که قادر است عبارت‌های مفهومی را از عبارت‌های متداول جدا کند. این ویژگی عبارت‌های مهم را در متن‌های مختلف شناسایی می‌کند.

۲-۴-۴- ویژگی‌های n-gram

n-gram ها قادر هستند مفهوم را تاحدودی حفظ کنند. این ویژگی به طور گسترده در پردازش زبان طبیعی کاربرد دارد و به دوصورت مبتنی بر کاراکتر و مبتنی بر کلمه وجود دارد. مراجع [۸, ۹, ۲۳,

۲۴] با مقایسه n-gram ها و ترکیب آن‌ها به این نتیجه رسیدند که (n > ۲) تاثیر جداسازی کمتری نسبت به کلمه‌ها دارند.

۲-۴-۵- منفی کننده‌ها[۱]

ارزش تشخیص منفی کننده‌ها در پردازش پاسخ نسبت به بازیابی اطلاعات سنتی، بیشتر است. وجود یک منفی کننده ساده در یک جمله به طوری تاثیرگذار است که می‌تواند نوع جمله را از مثبت به منفی تغییر دهد. در بعضی موارد وجود منفی کننده‌ها نوع جمله را تغییر نمی‌دهند. به طور مثال «این کتاب یکی از بهترین‌هاست هیچ شکی در آن نیست».

۲-۵- شناسایی و برچسب زدن اجزای واژگان زبان[۲]

شناسایی و برچسب زدن اجزای واژگان زبان، به معنای استفاده از یک تجزیه‌گر نحوی برای تعیین کاربرد کلمه به صورت فعل، اسم، صفت، قید، مضاف و مضاف علیه، ضمیر و غیره است [۲۵]. جدول (۲-۳) برخی از برچسب‌های واژگان را نمایش می‌دهد.

جدول (۲-۳): برچسب‌های واژگان در اجزای واژگان زبان

برچسب	توضیح	مثال
JJ	صفت	new ،good
JJR	صفت مقایسه‌ای	better
JJS	صفت برتر	wildest
FW	کلمه ناشناخته	Meta culpa
NN	اسم	family ،army
RB	قید	quickly ،never
RBR	قید مقایسه‌ای	faster
RBS	قید برتر	fastest
VB	فعل	left ،eat
PRP	ضمیر	he ،you ،I
IN	حرف اضافه	by ،in ،of
DT	مشخص کننده	the ،a

[۱] Negation

[۲] Parts Of Speech (pos)

نحوه استفاده از این تجزیه‌گر برای انتخاب کلمه‌های مناسب برای مجموعه مثبت و منفی به صورت زیر است:

۲-۵-۱- جداسازی صفت‌ها

صفت‌ها به عنوان ویژگی‌های پرتکرار در تمامی واژگان کلام به کار می‌روند. کاربران معمولا از صفت‌ها برای به تصویر کشیدن احساس خود استفاده می‌کنند. در [۹] تنها از وجود صفت‌ها برای کلاس‌بندی قطبی بر روی دیدگاه‌های توییتر انجام شده و محققان توانستند نتایجی با دقت بسیار بالایی توسط صفت‌ها بدست آورند.

۲-۵-۲-ترکیب صفت‌ها و قیدها

بیشتر قیدها به تنهایی بار مثبت و منفی ندارند. اما هنگامی که در کنار یک صفت قرار می‌گیرند، می‌توانند نقشی اصلی در تعیین درجه بار یک جمله داشته باشند. قیدها براساس توانایی آن‌ها در تعیین بار مثبت و منفی جملات به دسته‌های زیر تقسیم می‌شوند [۳]:

- قیدهای تأکیدی: قطعا، بدون شک و غیره
- قیدهای تردیدی: شاید، احتمالا و غیره
- قیدهای شدت، به شدت، بسیار وغیره
- قیدهای ضعیف: به ندرت، به سختی و غیره
- قیدهای منفی: هرگز، هیچگاه و غیره

۲-۶- جهت معنایی عبارات

یک عبارت ممکن است در پاسخ‌های هر سه دسته مثبت، منفی و خنثی تکرار شود؛ ولی بسته به شرایط و قرارگیری در کنار عبارت‌های خاص، ممکن است قطب متفاوتی پیدا کند [۲۶]. همنشینی[۱] کلمه‌ها به این معناست که یک کلمه به همراه چه کلماتی در یک متن ظاهر می‌شود. استفاده از

[۱] Collocation

معیارهایی که بتواند بار یک عبارت ناشناخته را نسبت به نوع دسته آن مشخص کند، برای دسته‌بندی مفید است. منبع [۲۶] جهت معنایی توسط معادله‌های (۲-۱) تا (۲-۳) محاسبه می‌کند.

$$Score^{+}(w_i) = \sum_{j=1}^{N^{+}} AssociationFunction(w_i, SeedPos_j) \quad (۲-۱)$$

$$Score^{-}(w_i) = \sum_{j=1}^{N^{-}} AssociationFunction(w_i, SeedNeg_j) \quad (۲-۲)$$

$$TotalScore(w_i) = Score^{+}(w_i) - Score^{-}(w_i) \quad (۲-۳)$$

$Score^{+}(w_i)$ امتیاز مثبت و $Score^{-}(w_i)$ امتیاز منفیعبارت i است. برای بدست آوردن امتیاز مثبت، از یک مجموعه عبارت آماده استفاده می‌شود که به عنوان عبارت‌های مثبت، در نظر گرفته می‌شوند. این مجموعه حاوی N^{+} عبارت مثبت است. ارتباط هر عبارت مثبت با عبارتی که بار آن مشخص نیست، توسط یک تابع نسبی محاسبه می‌شود. سپس امتیاز مثبت از مجموع ارتباط تمام عبارت‌های مثبت و عبارت ناشناخته بدست می‌آید. برای محاسبه امتیاز منفی نیز از یک مجموعه عبارت منفی با اندازه N^{-}، استفاده می‌شود. تمامی عملیات برای بدست آوردن این امتیاز مشابه امتیاز مثبت است. امتیاز کلی عبارت، از اختلاف امتیاز مثبت و امتیاز منفی بدست می‌آید. مقدار مثبت نشان‌دهنده مثبت بودن عبارت و مقدار منفی نشان‌دهنده منفی بودن آن است. هر چه این امتیاز به صفر نزدیک‌تر باشد، ارزش کلمه خنثی می‌شود.

۲-۶-۱- اطلاعات متقابل نقطه به نقطه[۱]

این تابع به صورت گسترده برای امتیازدهی بین دو عبارت در جهت معنایی استفاده می‌شود [۳, ۲۶]. اطلاعات متقابل نقطه به نقطه بین دو عبارت $term_1$ و $term_2$ از طریق فرمول (۲-۴) تعریف می‌شود[۲۷]:

$$PMI(term_1, term_2) = \log_2\left(\frac{p(term_1 \wedge term_2)}{p(term_1) \times p(term_2)}\right) \quad (۲-۴)$$

[۱] Point Wise Mut ual Information(PMI)

$p(term_i)$ احتمال رخداد عبارت i است. $p(term_1 \wedge term_2)$ احتمال این است که دو عبارت term_1 و term_2 با یکدیگر رخ دهند. اگر دو عبارت مذکور مستقل باشند احتمال آنها توسط $p(term_1) \times p(term_2)$ محاسبه می‌شود. لگاریتم تقسیم دو احتمال $p(term_1 \wedge term_2)$ و $p(term_1) \times p(term_2)$ بر یکدیگر، اطلاعاتی در مورد حضور یکی از عبارت‌ها هنگامی که عبارت دیگر مشاهده شود، بدست می‌آورد [۲۸]. ارتباط بین دو عبارت به صورت زیر تعیین می‌شود:

- اگر بین دو عبارت یک ارتباط واقعی وجود داشته باشد $PMI(term_1, term_2) \succ 0$
- اگر term_1 و term_2 مستقل باشند $PMI(term_1, term_2) = 0$
- اگر term_1 و term_2 مکمل یکدیگر باشند. $PMI(term_1, term_2) \prec 0$

۲-۷- روش‌های استخراج ویژگی[۱]

انتخاب زیر مجموعه مناسب از ویژگی‌ها در زمینه داده کاوی اهمیت زیادی دارد. این کار دقت طبقه-بندی را بهبود و زمان اجرای الگوریتم‌های یادگیری را کاهش می‌دهد[۱]. داده‌های با ابعاد بالا، باعث دشواری مراحل آموزش و تست روش‌های طبقه‌بندی عمومی می‌شوند. روش‌های مورد استفاده برای کاهش داده‌ها می‌تواند به دو روش بسته‌بندی[۲] و فیلتر[۳] تقسیم شوند [۲۹]. روش‌های بسته‌بندی به طور کلی عملکرد بهتری نسبت به روش‌های فیلتر دارد. با این حال هزینه پیچیدگی و زمان محاسباتی برای پایگاه داده‌هایی با ابعاد بزرگ در این روش‌ها گران است. رویکرد فیلتر در واقع قبل از فرایند طبقه‌بندی واقعی انجام می‌شود. روش فیلتر مستقل از الگوریتم یادگیری است، محاسبات آن ساده سریع و مقیاس‌پذیر است. با استفاده از روش فیلتر، انتخاب ویژگی یک بار انجام می‌شود و سپس می‌تواند به عنوان ورودی به طبقه‌بندی‌های داده شود.

برخی از روش‌های فیلتر، تنها ویژگی‌ها را رتبه‌بندی می‌کنند و عمل کاهش ویژگی را انجام نمی‌دهند. بنابراین آنها زمانی که نیاز به پیدا کردن تعداد مناسب از صفات دارند، با روش‌های جستجو ترکیب

---

[۱] Feature Extraction

[۲] Wrapper

[۳] Filter

می‌شوند. چنین فیلترهایی اغلب با انتخاب مستقیم[۱] مورد استفاده قرار می‌گیرند. به عنوان مثال روش‌های جستجو اول بهترین[۲]، جستجو دوطرفه[۳]، جستجو ژنتیک[۴]، حذف بازگشتی[۵] و غیره.

در بعضی از تحقیق‌ها مانند [۲۶, ۲۸]، از روش‌های استخراج ویژگی برای تعیین جهت معنایی عبارت-ها استفاده کرده‌اند. به طور کلی روش‌های استخراج ویژگی مشهور در دسته‌بندی قطبی عبارتند از:

۲-۷-۱- اطلاعات متقابل[۶]

در تحقیق‌های گذشته از اطلاعات متقابل به عنوان یک معیار برای همپوشانی بین دو متغیر تصادفی استفاده شده است. این تابع وظیفه‌های مختلفی مانند استخراج ویژگی [۳۰] و تشخیص همنشینی کلمه‌ها [۳۱] را انجام می‌دهد. در استخراج ویژگی، اطلاعات متقابل اندازه می‌گیرد چقدر اطلاعات حضور یا عدم حضور یک عبارت خاص، به تصمیم‌گیری صحیح در مورد یک کلاس، کمک می‌کند. نحوه محاسبه اطلاعات متقابل به صورت زیر:

$$MI(t,c) = \sum_{t\in\{0,1\}} \sum_{c\in\{0,1\}} p(t \wedge c) \times \log_2 \left(\frac{p(t \wedge c)}{p(t) \times p(c)} \right) \qquad (۵-۲)$$

t نماد ویژگی مورد بررسی است؛ c دسته‌ای است که بایستی در مورد آن تصمیم گرفت؛ $p(t \wedge c)$ امکان وقوع یا عدم وقوع عبارت t را در دسته c نشان می‌دهد؛ $p(t)$ احتمال رخداد یا عدم رخداد عبارت t را در کلیه دسته‌ها محاسبه می‌کند؛ $p(c)$ احتمال رخداد یا عدم رخداد دسته c را محاسبه می‌کند. در صورتی که c و t مستقل از یکدیگر باشند، $p(t \wedge c) = p(t) \times p(c)$ که باعث می‌شود مقدار اطلاعات متقابل برابر صفر گردد. تفاوت اطلاعات متقابل با اطلاعات متقابل نقطه به نقطه در این است که مقدار اطلاعات متقابل همواره بزرگتر از صفر است. در بعضی از مراجع مانند [۳۲] اطلاعات متقابل واطلاعات متقابل نقطه به نقطه را یکسان در نظر گرفته‌اند.

---

[۱] Forward Selection
[۲] Best-First Search
[۳] Bi-Directional
[۴] Genetic Search
[۵] Backward Elimination
[۶] Mutual Information(MI)

۲-۷-۳- بهره اطلاعات[۱]

مراجع [۱, ۳۲] بهره اطلاعات را یک روش مفید در استخراج ویژگی معرفی کردند. این روش، تعداد بیت‌های اطلاعاتی را که برای پیش‌بینی گروه‌ها به دست می‌آید را با بررسی وجود یا عدم وجود یک عبارت در یک سند، محاسبه می‌کند [۳۰]. فرمول (۲- ۶) نحوه محاسبه بهره اطلاعات را نمایش می-دهد.

$$IG(t) = -\sum_{i=0}^{|c|} P(c_i)\log P(c_i) + P(t)\sum_{i=0}^{|c|} P(c_i \mid t)\log P(c_i \mid t) + P(\bar{t})\sum_{i=0}^{|c|} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t}) \quad (۶-۲)$$

$P(c_i)$ احتمال اینکه دسته c_i رخ دهد؛ احتمال اینکه عبارت t رخ دهد؛ $P(\bar{t})$ احتمال اینکه عبارت t مشاهده نشود؛ $P(c_i \mid t)$ احتمال اینکه عبارت t در دسته c_i رخ دهد و $P(c_i \mid \bar{t})$ احتمال اینکه عبارت t در دسته c_i مشاهده نشود.

۲-۷-۴- نرخ بهره[۲]

نرخ بهره یک نسخه توسعه یافته از بهره اطلاعات است [۲۹] که توسط فرمول (۲-۷) محاسبه می‌شود:

$$GR(t) = \frac{IG(t)}{\sum_{i=1}^{m} P(t)\log P(t)} \quad (۷-۲)$$

$IG(t)$ بهره اطلاعات عبارت t است و $\sum_{i=1}^{m} P(t)\log P(t)$ آنتروپی عبارت t را محاسبه می‌کند. i دسته را نشان می‌دهد. مرجع [۱] با استفاده از این روش دقت بالاتری نسبت به بهره اطلاعات بدست آورد. یک عیب اساسی در نرخ بهره این است که ممکن است مخرج صفر شود و یا خیلی کوچک شود. برای حل این مشکل از روش‌های هیورستیک استفاده می‌شود [۲۹].

---

۱ Information Gain(GI)

۲ Gain Ratio (GR)

۲-۷-۵- مربع کای[۱]

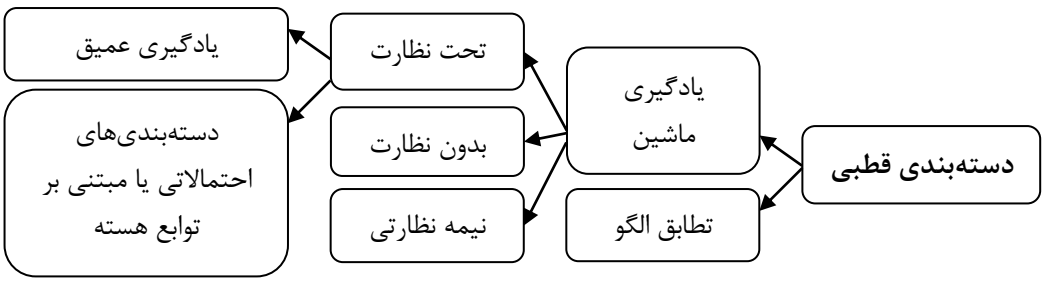
همانند اطلاعات متقابل، این تابع نیز وظایف مختلفی را در پردازش متن انجام می‌دهد ولی در زمینه کلاس‌بندی قطبی، بیشتر برای استخراج ویژگی استفاده می‌شود [۳۳]. برای هدف استخراج ویژگی، میزان وابستگی حضور یک عبارت و یک دسته خاص از طریق مربع کای محاسبه می‌گردد. روش بدست آوردن مربع کای بین عبارت t و دسته c در متن D از طریق فرمول (۲-۸) انجام می‌پذیرد [۳۴]:

$$\chi^2(D,t,c) = \sum_{t\in\{0,1\}} \sum_{c\in\{0,1\}} \frac{(O_{tc} - E_{tc})}{E_{tc}} \qquad (۲-۸)$$

O_{tc} تعداد تکرار واقعی رخداد یا عدم رخداد عبارت t در دسته c است؛ E_{tc} تعداد تکرار مورد انتظار رخداد یا عدم رخداد عبارت t در دسته c است. میزان مربع کای بالا به معنای وابسته بودن عبارت در دسته مورد نظر است. بزرگترین عیب مربع کای این است که در صورتی که متن کوتاه و خلوت باشد، نمی‌تواند به خوبی ارتباط بین دو عبارت را محاسبه کند[۳۴].

۲-۸- روش‌های مشهور در دسته‌بندی قطبی

پس از انجام عملیات مربوط به پیدا کردن مشخصه‌های مناسب برای نمایش پاسخ مورد نظر، بایستی روشی مناسب برای دسته‌بندی پاسخ اتخاذ گردد. تاکنون روش‌های مختلفی برای دسته‌بندی قطبی متون مختلف آزمایش شده‌اند. منبع [۳۵] روش‌های مختلف دسته‌بندی را بررسی کرده و به صورت زیر معرفی می‌کند.



شکل (۲-۱): روش‌های دسته‌بندی قطبی

[۱] Chi Square

۲-۸-۱- روش‌های مبتنی بر یادگیری ماشین

رویکرد یادگیری ماشین در پردازش متن بسیار کاربردی هستند. زیرا کامل و به صورت خودکار اجرا می‌شوند و می‌توانند با مجموعه داده‌های بزرگ کار کنند. روش‌های یادگیری ماشین را می‌توان به سه دسته روش‌های یادگیری تحت نظارت، بدون نظارت و نیمه نظارت تقسیم کرد.

یادگیری تحت نظارت: یادگیری تحت نظارت یک راه حل بالغ و موفق در طبقه‌بندی موضوعی سنتی است که به طور گسترده در دسته‌بندی قطبی استفاده می‌شود [۳۵]. برای هر دو طبقه بندی موضوعی و قطبی، **بیزین ساده**[۱] و **ماشین بردار پشتیبان**[۲] رایج‌ترین و موثرترین الگوریتم‌های یادگیری تحت نظارت هستند [۱, ۹, ۱۰, ۲۳, ۳۳, ۳۶].

بزرگترین محدودیت یادگیری نظارت شده این است که حساس به کمیت و کیفیت داده‌های آموزشی هستند. همچنین الگوریتم ممکن است در مواقعی که داده‌های آموزشی جانبدارانه و یا ناکافی باشند، شکست بخورد. دسته‌بندی پاسخ در سطح‌های جزئی‌تر، چالش‌های بیشتری را برای رویکردهای مبتنی بر یادگیری تحت نظارت ایجاد می‌کند. زیرا اطلاعات کمی برای طبقه‌بندی وجود دارد.

یادگیری بدون نظارت: در طبقه‌بندی متن، گاهی اوقات برچسب‌زنی داده‌ها و یا یافتن آن‌ها دشوار است. اما جمع‌آوری داده‌های بدون برچسب مشکل نیست. روش‌های یادگیری بدون نظارت برای دسته‌بندی اینگونه داده‌ها هستند.

محدودیت رویکردهای بدون نظارت در این است که به طور معمول به حجم زیادی از اطلاعات برای آموزش دقیق نیاز دارند. مدل‌های بدون نظارت اغلب موضوعات غیر منسجم ایجاد می‌کنند. زیرا توابع هدف آن‌ها همیشه با قضاوت انسانی همخوانی ندارند. با این حال، یادگیری بدون نظارت راهی برای بدست آوردن دانش در مورد داده‌ها، بدون نیاز به هیچ گونه برچسب‌زنی است.

---

[۱] Naïve Bayes

[۲] Support Vector Machine (SVM)

یادگیری نیمه نظارتی: مدل‌های یادگیری نیمه نظارتی می‌توانند با استفاده از هر دو نوع داده برچسب و بدون برچسب یاد بگیرند. انگیزه اصلی استفاده از مدل‌های نیمه نظارتی در دسته‌بندی قطبی، فقدان داده های برچسب‌گذاری شده در برنامه‌های دنیای واقعی است. با توجه به [۳۳]، ایده اصلی یادگیری نیمه نظارتی این است که اگر چه داده‌های بدون برچسب حاوی اطلاعاتی در مورد کلاس‌ها نیستند، اما آنها اطلاعاتی در مورد توزیع مشترک بر روی ویژگی‌های طبقه‌بندی دارند. بنابراین، هنگامی که هر دو نوع داده برچسب‌دار و بدون برچسب موجود است، استفاده از یادگیری نیمه نظارتی می‌تواند بهبود بیشتری نسبت به یادگیری بدون نظارت داشته باشد. همچنین اگر به مدل‌های یادگیری بدون نظارت دانش قبلی اضافه شود، دیگر محدودیت‌های آن را نخواهد داشت.

۲-۸-۲-تطابق الگو

در بعضی از مطالعات از مجموعه عبارت‌های از پیش‌تعیین شده برای دسته‌بندی قطبی استفاده می-کنند [۳۷]. همچنین یکی از مسائل مهم در تعیین جهت معنایی عبارت‌های ناشناخته، تعیین مجموعه عبارت‌های مثبت و منفی مناسب است. تابع نسبی با استفاده از این مجموعه‌ها باید بتوانند امتیاز عبارت ناشناخته را به درستی محاسبه کند. به روش استفاده از الگوهای آماده در تجزیه و تحلیل متن، رویکرد مبتنی بر واژگان[۱] یا تطابق الگو می‌گویند. برای تطابق الگو دو روش وجود دارد. رویکرد مبتنی بر لغت‌نامه که وابسته به شناسایی کلمات احساسی است و پس از شناسایی، لغت‌های مترادف جستجو می‌شوند. رویکرد دوم مبتنی بر مجموعه نوشته است که با یک لیست از کلمات احساسی آغاز می‌شود و سپس کلمات احساسی دیگر در یک متن بزرگ پیدا می‌شوند. این رویکرد به پیدا کردن کلمات نظر با جهت خاص مفهومی کمک می‌کند [۳۵]. برخی از لغت‌نامه‌های مشهور در این زمینه عبارتند از:

[۱] Lexicon-Based Approach

لغت‌نامه احساسی NRC[1]: این لغت‌نامه با استفاده از سیستم ترجمه گوگل[2] تکمیل شده و حاوی لغت‌های احساسی ۲۰ زبان مختلف است [۳۸]. هرکلمه در این لغت‌نامه علاوه مشخص بودن نوع قطب مثبت و منفی بودن آن، میزان عواطف مختلفی که در معنی آن وجود دارد مانند خشم، ترس، شادی، غم و اطمینان نیز نمایش داده می‌شود.

WordNet: WordNet یک لغت‌نامه گسترده انگلیسی است که در سال ۱۹۸۵ توسط دانشگاه پرینستون ارائه گردید. در این لغت‌نامه اسم‌ها، فعل‌ها، صفت‌ها و قیدها به مجموعه‌ای از مترادف‌ها گروه‌بندی می‌شوند[۳۹]. این لغت‌نامه به صورت رایگان برای دانلود در دسترس است. ساختار WordNet یک ابزار مفید برای زبان‌شناسی محاسباتی و پردازش زبان طبیعی است. در حال حاضر این لغت‌نامه برای زبان‌های مختلفی مانند عربی، فرانسوی، ترکی، بلغاری، چینی، ایتالیایی، آلمانی، یونانی و غیره موجود است.

SENTIWORDNET: مرجع [۴۰] یک منبع واژگانی به نام SENTIWORDNET 3.0 را برای پشتیبانی از طبقه‌بندی احساس و پردازش دیدگاه، ارائه می‌دهد. این لغت‌نامه از واژگان WordNet استفاده می‌کند و کلمه‌ها را با توجه به درجه مثبت، منفی و خنثی بودن آن‌ها، برچسب می‌زند. این لغت‌نامه برای اهداف آنالیز احساسی بوجود آمده است.

۲-۹- جمع‌بندی

در این فصل روش‌های مشهور و موثر در دسته‌بندی قطبی بررسی شدند. در ابتدای دسته‌بندی، نیازمند انتخاب ویژگی‌های مفید و قدرتمند برای جداسازی دسته‌ها است. معمولا مجموعه ویژگی‌هایی که برای دسته‌بندی متن استفاده می‌شود ابعاد بزرگی دارد که باعث می‌شود زمان پردازش افزایش یابد. برای جلوگیری از این مشکل بایستی از روش‌های مناسب برای کاهش ابعاد استفاده کرد. پس از

[1] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.html

[2] translate.google.com

عملیات مربوط به انتخاب و کاهش ابعاد ویژگی‌ها، به بررسی روش‌ها و تکنیک‌هایی که برای دسته-بندی قطبی استفاده می‌شوند، پرداخته شد.

جدول (۲-۴): مقایسه روش‌های دسته‌بندی قطبی در منابع مختلف

شماره	منبع	پایگاه‌داده	پیش‌پردازش	ویژگی‌ها	روش	ارزیابی
۱	[۱]	پایگاه اینترنتی فیلم (IMDb) با ۱۰۰۰ نظر مثبت و ۱۰۰۰ نظر منفی	حذف کلمات اضافی و ریشه یابی	کلمه‌ها	با استفاده از روش‌های کاهش ویژگی فرکانس سند، بهره اطلاعات، مربع کای، نرخ بهره و باور F ویژگی‌های مناسب بین بازه (۱۰۰-۱۰۰۰) انتخاب شدند و با استفاده از روش‌های ماشین بردار پشتیبان (RBF)، بیزین ساده، درخت تصمیم، حداکثر آنتروپی، k نزدیک‌ترین همسایه، طبقه‌بند Winnow و Adaboost آزمایش شدند.	بهترین دقت توسط نرخ بهره و بیزین ساه بدست آمد. این دقت برابر ۹۰/۹۰٪ است. نرخ بهره اندازه ویژگی‌ها را به ۵۰۰ کاهش داد.
۲	[۸]	مسابقات مهم بین المللی SemEval (انگلیسی)، TASS (اسپانیایی) و SENTIPOLC (ایتالیایی)	حذف علائم نشانه-گذاری و لهجه، ریشه‌یابی، حذف کلمات اضافی و کاهش نماد	TF-IDF آزمایش شده بر روی n-gram های مبتنی بر کلمه و کاراکتر	ماشین بردار پشتیبان (خطی)	دقت در زبان انگلیسی ۶۲/۹، در زبان اسپانیایی ۶۳/۷ و در زبان ایتالیایی ۶۱ بدست آمد.
۳	[۹]	پایگاه اینترنتی فیلم (IMDb) با ۱۰۰۰ نظر مثبت و ۱۰۰۰ نظر منفی	-	استفاده از دو عامل انسانی برای پیدا کردن کلمه‌های احساسی؛ فرکانس، حضور و موقعیت عبارت بر روی unigram ها و bigram ها؛ برچسب‌گذاری اجزای واژگان زبان،	آزمایش بیزین ساده، حداکثر آنتروپی، و ماشین بردار پشتیبان	Unigram های مبتنی بر حضور عبارت با دقت ۸۲/۹ بهترین دقت را بدست آورد. تعداد ویژگی‌ها ۱۶۱۶۵ unigram است.

شماره	منبع	پایگاه‌داده	پیش‌پردازش	ویژگی‌ها	روش	ارزیابی
۴	[۱۰]	نتایج انتخاب آمریکا سال ۲۰۱۲ انتخابات هند سال ۲۰۱۳	تبدیل حالت‌های نامتداول و اختصار به فرمت مناسب؛ حذف لینک‌ها و تصویر؛ حذف کلمات تکراری پی در پی؛ حذف تکرار نماد	فرکانس عبارت بر روی unigram، bigram و ترکیب آن‌ها	ماشین بردار پشتیبان، بیزین ساده، ماکزیمم آنتروپی و شبکه عصبی؛ جهت کاهش ویژگی از تحلیل مولفه اصلی استفاده شده است	بهترین دقت توسط ماشین بردار پشتیبان و تحلیل مولفه اصلی با مقدار ۹۳٪ بدست آمد.
۵	[۱۱]	۳۹۰۲ عقیده هتل	تجزیه کلمه‌ها توسط WordNet	برچسب‌گذاری اشکال واژگان سخن	ماشین بردار پشتیبان و دامنه آنتولوژی فازی	دقت برای اتاق ۷۰/۹٪، برای رستوران ۷۱٪، برای موقعیت ۷۲/۷٪، برای خدمات ۷۳/۷٪، برای کارکنان ۷۲/۹٪ و برای استخر و جکوزی ۷۰٪ بدست آمد.
۶	[۲۲]	پایگاه اینترنتی فیلم (IMDb) با ۱۰۰۰ نظر مثبت و ۱۰۰۰ نظر منفی؛ نظرهای آمازون در مورد کتاب، DVD، وسایل آشپزخانه و وسایل الکترونیکی	حذف نشانه‌های نگارشی، تبدیل حروف و حذف کلماتی که بیشترین یا کمترین تکرار را داشتند	TF-IDF	ماشین بردار پشتیبان (خطی)	بهترین دقت توسط فرکانس عبارت BM25 و دلتا BM25 به عنوان فرکانس معکوس سند با مقدار ۹۶/۹۰٪ بدست آمد

شماره	منبع	پایگاه‌داده	پیش‌پردازش	ویژگی‌ها	روش	ارزیابی
۷	[۲۳]	پایگاه اینترنتی فیلم (IMDb) با ۱۰۰۰ نظر مثبت و ۱۰۰۰ نظر منفی	حذف کلمات اضافی و اعداد	آزمایش TF-IDF بر روی unigram ها، bigram ها، trigram ها و ترکیب آنها	بیزین ساده، ماکزیمم آنتروپی، کاهش گرادیان نزولی و ماشین بردار پشتیبان	بهترین دقت توسط ماشین بردار پشتیبان و unigram ها با مقدار ۸۸/۹۴ بدست آمد.
۸	[۲۶]	۵۰۰ سند دیدگاه مدل‌های مختلف تلفن همراه در آمازون	-	TF-IDF	از روش‌های کاهش ویژگی اطلاعات متقابل نقطه به نقطه، اطلاعات متقابل، بهره اطلاعات، مربع کای و نرخ شباهت لگاریتی برای کاهش ویژگی استفاده گردید. این روش‌ها برای پیدا کردن جهت معنایی کلمه‌ها استفاده شدند. از بیزین ساده، درخت تصمیم، شبکه عصبی چند لایه پرسپترون و الگوریتم bagging (REP Tree) استفاده شد.	درخت تصمیم بهترین معیار-f را بدست آورد این مقدار برابر است با ۰/۸۵۵
۹	[۲۸]	۴۱۰ دیدگاه در حوزه‌های خودرو، فیلم و مسافرت	-	صفت‌ها	جهت‌گیری عبارت‌ها توسط اطلاعات متقابل نقطه به نقطه	دقت برای دیدگاه‌های خودرو ۸۴٪، ۶۶٪ فیلم و ۷۴٪ مسافرت بدست آمد.

فصل سوم:

سیستم‌های پرسش و پاسخ تعاملی

۳-۱- مقدمه

سیستم پرسش و پاسخ تعاملی، زمینه تحقیقاتی جدیدی است که در تقاطع سیستم‌های پرسش و پاسخ و سیستم‌های گفتگو به وجود آمده‌اند. در این سیستم، کاربران می‌توانند با استفاده از زبان‌های طبیعی پرسش‌های خود را مطرح کنند و در صورت امکان، پاسخ واقعی را به سوال ارجاع دهند. هنگامی که پاسخ سوال کاربر بیش از اندازه کم یا زیاد باشد و یا درخواست مبهم باشد، گفتگو می‌تواند حالت تعاملی داشته باشد. علاوه بر این در طی فرایند پاسخ‌گویی، سیستم می‌تواند گفتگو با کاربر را برای روشن شدن اطلاعات پیچیده یا مبهم آغاز کند یا موضوعات بیشتری را برای گفتگو پیشنهاد دهد. همچنین به کاربران اجازه می دهد که سوالات اضافی را در صورتی که نتایج به دست آمده مورد نیاز آن‌ها نبوده و یا در مواردی که به اطلاعات بیشتری نیاز دارند را بپرسند. سیستم پرسش و پاسخ تعاملی این ویژگی‌ها را از سیستم‌های پرسش و پاسخ و سیستم‌های گفتگو به ارث می‌برد. بنابراین در این فصل علاوه بر معرفی سیستم‌های پرسش و پاسخ تعاملی، این سیستم‌ها نیز معرفی می‌شوند. همچنین نحوه تعامل با کاربر نیز در این فصل مطالعه می‌شود.

۳-۲- خصوصیات گفتگوی انسانی

یک مکالمه از گروهی از افراد تشکیل شده است که در مورد موضوعات خاصی با یکدیگر گفتگو می‌کنند. بدلیل این که امروزه تعاملات اجتماعی بیشتر در اینترنت اتفاق می‌افتد، محققان علاقه‌مند به مطالعه مکالمه در رسانه‌های اجتماعی آنلاین شده‌اند. از طریق این مطالعه‌ها، محققان می‌توانند چگونگی انجام گفتگوهای انسانی و بحث‌ها را بررسی کنند [۴۱]. مکالمه انسانی یک فعالیت دشوار است که شامل بسیاری از فرآیندهای مختلف از جمله درک زبان طبیعی و تولید واحدهای گفتگو می‌شود. زبان و تعاملات در هر مکالمه و هر گروه پیچیده و متغیر است. چهار ویژگی نوبت گرفتن، زمینه، ساختار و رهبری مکالمه در مکالمه‌های انسانی وجود دارند که در هنگام ایجاد سیستم گفتگو و یا

پرسش و پاسخ تعاملی باید مورد توجه قرار گیرند. همه این خصوصیات گفتگوها، به عنوان تجزیه و تحلیل مکالمه[1] در [۴۲] بررسی شده است که در ادامه هر یک توضیح داده می‌شوند.

۳-۲-۱- نوبت گرفتن

مکالمه انسانی بوسیله نوبت‌گیری در یک ساختار متوالی شکل می‌گیرد، به این صورت که یک نفر صحبت را آغاز می‌کند و پس از آن، شخص دیگری نوبت خود را شروع می‌کند. این خصوصیت تعیین می‌کند تا چه میزان سیستم طبیعی رفتار می‌کند. معمولا انسان‌ها به راحتی تشخیص می‌دهند که در چه زمانی گفتگو را آغاز کنند، اما برای سیستم‌های گفتگو، نوبت‌گیری می‌تواند مشکل باشد. زیرا در هنگام نوبت‌گیری، سکوت طولانی یا همپوشانی دو نوبت، مکالمه را از حالت طبیعی خارج می‌کند[۴۳]. با این حال حتی در مکالمات انسانی، کمتر از ۱٪ از تغییرات نوبت، فاصله صفر واقعی (سکوت کمتر از ۱۰ میلی ثانیه) و بیش از ۴۰٪ از انتقال‌ها همپوشانی دارند[۴۴]. واضح‌ترین راه این است که تصور شود هنگامی که سکوت آغاز می‌شود، نوبت‌گیری باید انجام شود. اما مطالعات نشان می‌دهند که سکوت در نوبت‌ها بسیار کوچک است و انسان‌ها حتی زمانی که سخنران قبلی صحبت خود را متوقف می‌کند به راحتی می توانند درک کنند که نوبت کدام یک از آنهاست. به همین دلیل، درک گذشته گفتگو برای پیش‌بینی نوبت بعدی بسیار مهم است.

۳-۲-۲-زمینه

دومین مشخصه‌ای که باید در هنگام ایجاد سیستم‌های گفتگو مورد توجه قرار گیرد، زمینه است. زمینه به معنای موضوع بحث است. به منظور برقراری ارتباط موفق، افراد باید اطلاعاتی داشته باشند که به عنوان پس‌زمینه برای تفسیر سخنان و نیت‌های کسی که نوبت را در دست دارد، استفاده کنند. با شروع از یک زمینه مشترک، شرکت‌کنندگان با اضافه کردن مقدار کمی اطلاعات، یک گفتگوی منسجم را می‌سازند. با این حال، شرکت‌کنندگان در یک مکالمه، پیش‌فرض خود را دارند و این مهم است که اطمینان حاصل شود که آنها یکدیگر را درک کنند. گاهی اوقات با استفاده از عباراتی مانند

[1] Conversational Analysis

«درسته»، «بله» یا از طریق سؤالات اضافی، توسط مخاطبان اعلام می‌شود که گفتگو را درک می-کنند. همه اینها تضمین می‌کند که مکالمه حاصل از آن معنی‌دار است.

۳-۲-۳- ساختار مکالمه

یکی دیگر از خصوصیات مکالمه، ساختار خاص آن است. ساختار مکالمه شامل اطلاعاتی برای برقراری ارتباط، محتوای نوبت، تعداد و نقش هر مخاطب است. کشف ساختار برای شناسایی اهداف اجتماعی تعاملات انسانی در مکالمه‌ها و در بسیاری از برنامه های کاربردی از جمله خلاصه سازی [۴۵]، سیستم‌های گفتگو [۴۶] و بازی‌های گفتگو [۴۷] مهم است. انسان‌ها دانش ضمنی از ساختار مکالمه دارند و می‌توانند به راحتی مکالمات طبیعی را از مصنوعی جدا کنند. هنگام تعریف یک واحد تحلیل برای ساختار مکالمه، از اندازه‌های مختلفی استفاده می‌شود. به عنوان مثال، برچسب‌گذاری واژگان سخن تنها بر روی کلمات در توالی‌ها انجام می‌شود و درختان تجزیه‌گر زبان در سطح جمله هستند. بررسی ساختار مکالمه می‌تواند در سطح‌های مختلفی از جمله واژگان، جمله، سند و یا هر طول دلخواه از متن انجام شود[۴۸]. در حالی که الگوهای لفظی در یک ساختار درختی در سطح جمله یا پاراگراف تعریف می‌شوند. در مکالمه، شایع‌ترین واحد تجزیه و تحلیل، سخن[۱] است. طول این واحد بین سطح کلمه و جمله متغیر است. تجزیه تحلیل سخن در سیستم‌های گفتگو معمولا از طریق اعلان مکالمه[۲] انجام می‌شود که در بخش (۳-۹-۲) توضیح داده می‌شود.

۳-۲-۴-رهبری مکالمه[۳]

یکی دیگر از ویژگی‌های مهم گفتگو، مفهوم ابتکار عمل است که نشان می‌دهد چه کسی مکالمه را کنترل می‌کند و رهبری مکالمه را در دست دارد. به طور کلی سه نوع استراتژی کنترل، در سیستم-های گفتگو وجود دارد. این سه روش عبارتند از: کنترل توسط کاربر، کنترل توسط سیستم و کنترل ترکیبی[۴۹]. در سیستم‌هایی که کاربر کنترل مکالمه را در دست دارد، سیستم تنها می‌تواند به

---

[۱] Utterance

[۲] Dialogue Act

[۳] Initiative

درخواست‌ها و دستورات کاربر پاسخ دهد. مشکل اصلی این استراتژی این است که کاربر می‌تواند هر چیزی که می‌خواهد بپرسد و از مسیر مکالمه خارج شود. سیستم گفتگوی EDECAN مثالی از این نوع سیستم است که برای رزرو تجهیزات ورزشی در دانشگاه فنی والنسیا اسپانیا استفاده می‌شود [۵۰]. هنگامی که تنها سیستم، گفتگو را هدایت می‌کند، کاربر فقط می‌تواند به پرسش‌های سیستم، پاسخ دهد. مزیت این استراتژی این است که به محدود کردن ورودی کاربر کمک می‌کند و منجر به گفتگوهای کارآمدتر می‌شود. عیب آن کمبود انعطاف‌پذیری است، زیرا کاربر با توجه به انتظارات سیستم رفتار می‌کند. ربات چت WOSBOT مثالی از این نوع سیستم است که هدف آن رزرو هتل و پرواز است[۵۱]. در کنترل ترکیبی، کاربر و سیستم هر دو می‌توانند مکالمه را رهبری کنند. مزیت این روش این است که کاربر می‌تواند سیستم را در وظایفی که باید انجام شود، هدایت کند. در حالی که سیستم می‌تواند سؤال‌ها و موضوعات جدید را مطرح کند و پاسخ‌های بیش‌تری را ارائه دهد. سیستم گفتگو توسعه‌یافته LUNA مثالی از این نوع سیستم است اطلاعات راه آهن مانند جدول زمانی، هزینه ها، زمان سفر، نوع قطار و خدمات کاربر و غیره را فراهم می‌کند[۵۲].

۳-۳- انواع مکالمه تحت وب

ساختار مکالمه در گفتگوهای گفتاری و نوشتاری با یکدیگر متفاوت است. در مکالمات گفتاری، گفتگو بین گویندگان هماهنگ شده است. گویندگان ایده‌های یکدیگر را می‌شنوند و سپس نظرات خود را بیان می‌کنند. بنابراین ترتیب زمانی[۱] می‌تواند به عنوان ساختار مکالمه در این نوع مکالمات در نظر گرفته شود. از این رو به این گونه گفتگوها، مکالمه همزمان[۲] گفته می‌شود. مکالمات اسکایپ[۳] و کلاس‌های درس آنلاین، نمونه‌ای از این نوع مکالمه هستند. با این حال، در مکالمات نوشتاری مانند ایمیل‌ها و انجمن‌ها، نویسندگان در ترتیب‌های متفاوت بحث می‌کنند و گاهی اوقات به محتوای پست‌های قبلی توجه نمی‌کنند. بنابراین، ترتیب زمانی مکالمه نمی‌تواند به عنوان ساختار گفتگو در

[۱] Temporal Order

[۲] Synchronous Conversations

[۳] Skype

این حوزه مورد استفاده قرار گیرد و باید از تکنیک‌های مناسب برای استخراج ساختار زیربنایی در این مکالمات استفاده شود [۴۶]. اینگونه گفتگوها، معمولا مکالمات غیر همزمان[۱] نامیده می‌شوند.

۳-۴- سیستم‌های گفتگو

سیستم‌های گفتگو به سیستم‌هایی گفته می‌شود که بتواند با انسان‌ها و یا سیستم‌های دیگر یک گفتگوی منسجم انجام دهد. این گفتگو از طریق زبان انسانی و معمولا در چند نوبت انجام می‌شود [۵۳].

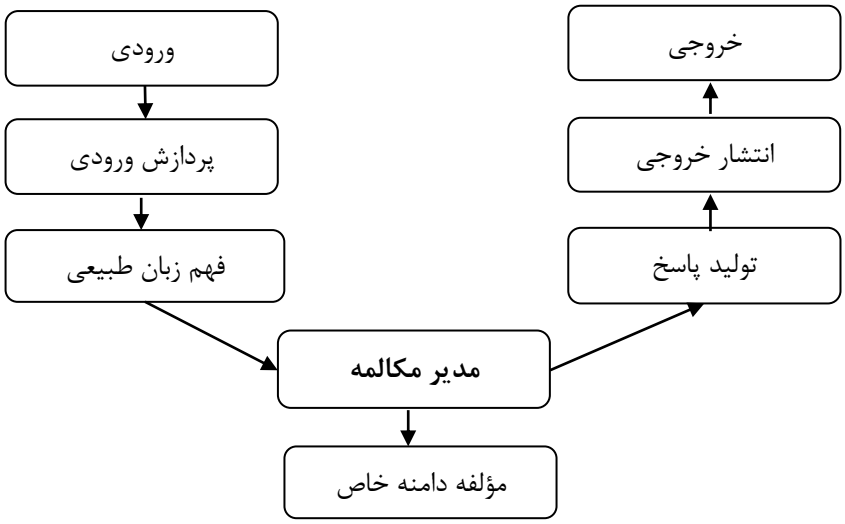
این تعریف چندین جنبه مهم از چنین سیستم‌هایی را برجسته می‌کند. یک سیستم گفتگو همیشه یک مخاطب دارد که برای هدف خاصی مانند تکمیل برخی کارها با سیستم تعامل می‌کند. تعامل شامل گفتگو در زبان انسانی بین دو یا چند شرکت‌کننده است و در چند نوبت انجام شود. تعامل از طریق زبان انسانی انجام می‌شود. کاربر نیازی به دانستن زبان‌های رایانه‌ای مانند SQL و XML ندارد. این سیستم‌ها به فرمت‌های متنی، گرافیکی، گفتاری و ترکیبی در حال توسعه هستند و می توانند در تلفن‌های هوشمند، سیستم‌های PDA، اتومبیل‌ها، سیستم‌های ربات و مرورگرهای وب مورد استفاده قرار گیرند. این سیستم‌ها به دو دسته هدفمند[۲] و بدون هدف[۳] تقسیم می‌شوند. سیستم‌های گفتگوی هدفمند کاربر را برای تکمیل کار خاصی، از طریق گفتگوها هدایت می‌کند و آموزش آن نیازمند بازخورد کاربر در مورد کیفیت گفتگو است. هدف ضمنی یک سیستم گفتگو بدون هدف با دامنه باز، ایجاد یک مکالمه منسجم، دلپذیر و جذاب است که به ورودی کاربر پاسخ می‌دهد [۱۸]. مکالمه‌های این نوع سیستم گفتگو مانند مکالمه روزمره است.

۳-۵- معماری سیستم‌های گفتگو

سیستم‌های گفتگوی ابتدایی، معماری بسیار ساده‌ای داشتند و تنها بر تطبیق الگو و حضور کلمات کلیدی خاص در سخنان انسانی برای تولید یک خروجی تمرکز داشتند. چند نوبت اول گفتگو با یک

[۱] Asynchronous Conversations
[۲] Task-Oriented Dialog System
[۳] non-task-oriented dialog system

ربات چت ممکن است معمولی به نظر برسد، اما در نوبت‌های بعدی مکالمه به سرعت بی‌معنی می-شود. به همین دلیل، سیستم‌های گفتگوی فعلی پردازش پیچیده‌تر و ماژول‌های مختلفی دارند. ساختار سیستم‌های گفتگو نسبت به یکدیگر بسیار متفاوت است، اما در منبع [۵۴] محققان معتقدند که معمولا آنها شامل شش قسمت اصلی پردازش ورودی[۱]، درک زبان طبیعی[۲]، مدیر مکالمه[۳]، مؤلفه دامنه خاص[۴]، تولید پاسخ[۵] و انتشار خروجی[۶] هستند. این اجزا در شکل (۳-۱) نمایش داده شدند.



شکل (۳-۱): مولفه‌های اصلی یک سیستم گفتگو

۳-۵-۱- پردازش ورودی

این مولفه تنها در سیستم‌های گفتگویی وجود دارد که مبتنی بر متن نیستند و ورودی را به یک متن ساده تبدیل می‌کند. این جزء در سیستم‌های گفتگوی صوتی، نیاز به دانش فونتیک[۷] و واج‌شناسی[۸] دارد. فونتیک شاخه‌ای از زبان‌شناسی است که با تولید و ترکیبات اصوات و نمایش آنها با استفاده از نمادهای متنی سروکار دارد. واج‌شناسی به بررسی نظام آوایی زبان می‌پردازد و جایگاه عناصر آوایی را

[۱] Input Decoder
[۲] Natural Language Understanding(NLU)
[۳] Dialogue Management
[۴] Domain Specific Component
[۵] Response Generator
[۶] Output Renderer
[۷] Phonetics
[۸] Phonology

در نظام زبان مشخص می‌کند. سیستم‌های زیادی وظیفه این مولفه را انجام می‌دهند. این موارد عبارتند از تشخیص خودکار گفتار[۱]، تشخیص گفتار کامپیوتر[۲] یا به عبارت ساده‌تر، تبدیل گفتار به متن[۳]. علاوه بر گفتار، سیستم گفتگو می‌تواند براساس گرافیک، دست خط و غیره باشد.

۳-۵-۲- فهم زبان طبیعی

هدف این واحد این است که بتواند گفته‌های کاربر را درک کند. این واحد، توالی کلمات را به نمایش معنایی تبدیل می‌کند که مدیر گفتگو می‌تواند در مرحله بعدی از آن استفاده کند. برای ایجاد این مولفه از دو رویکرد دستی و داده محور استفاده می‌شود. این جزء از مورفولوژی، نحو و معنا استفاده می‌کند. مورفولوژی مطالعه ساختار و محتوای اشکال کلمه است. پس از شناسایی کلمات کلیدی و تشکیل یک معنی آن را به مدیر گفتگو ارائه می دهد.

۳-۵-۳- مدیر گفتگو

مدیر گفتگو یکی از مهمترین اجزای سیستم گفتگو است؛ زیرا فعالیت چندین زیرمجموعه را در یک سیستم گفتگو هماهنگ می‌کند. هدف اصلی آن حفظ نمایش وضعیت کنونی مکالمه است. منبع [۴۹] وظایف اصلی مدیر گفتگو را به شرح زیر بیان کرده است:

- به روزرسانی زمینه گفتگو بر اساس تفسیر ارتباطات؛
- حفظ سابقه گفتگو؛
- مقابله با متن ناقص و ناشناخته؛
- تصمیم‌گیری در مورد محتوای نوبت بعدی و زمان بیان آن؛
- بازیابی محتویات ذخیره شده در فایل‌ها یا پایگاه‌داده؛
- ارائه بهترین پاسخ به کاربر.

---

[۱] Automatic Speech Recognition (ASR)

[۲] Computer Speech Recognition

[۳] Speech to Text

بعضی از این توابع می‌تواند به مولفه‌های دیگر انتقال یابد. به طورکلی مدیر مکالمه باید اعلان‌های مکالمه را تفسیر کند، اقدامات حل مسئله را انجام دهد و وضعیت گفتگو را حفظ کند. بنابراین، مدیر گفتگو کلیه معماری و ساختار گفتگو را کنترل می‌کند و همچنین به عنوان یک واسط بین مؤلفه‌های فهم زبان طبیعی و تولید پاسخ عمل می‌کند، زیرا اطلاعات را از فهم زبان طبیعی بدست می‌آورد، پردازش‌های لازم را انجام می‌دهد و به ماژول تولید پاسخ منتقل می‌کند.

۳-۵-۴- مولفه دامنه خاص

مدیر گفتگو معمولا نیاز به ارتباط با برخی از نرم‌افزارهای خارجی مانند پایگاه داده و یا یک سیستم خبره دارد. بنابراین درخواست یا برنامه مورد نظر، باید از نمایش داخلی استفاده شده توسط مدیر گفتگو به قالب خاص آن نرم‌افزار(مثلا SQL) تبدیل شود. این واسط توسط مولفه دامنه خاص انجام می‌شود. این مولفه می‌تواند توسط سیستم پردازش زبان کوئری طبیعی[۱] مدیریت شود. این سیستم کوئری‌های SQL را از زبان طبیعی ایجاد می‌کند.

۳-۵-۵- مولفه تولید پاسخ

این مولفه در مورد اینکه چه اطلاعاتی باید به کاربر نمایش داده شود، چگونگی ساخت اطلاعات، انتخاب کلمات و ساختار نحوی پیام، تصمیم‌گیری می‌کند. سیستم‌های کنونی از روش‌های ساده مانند قرار دادن داده‌های بازیابی شده در فرمت‌های از پیش تعریف شده استفاده می‌کنند. در یک سیستم گفتگو، ورودی مولفه تولید پاسخ، خروجی مدیر گفتگو است. این بدان معنی است که مدیر گفتگو محتوای پاسخ را تعیین می‌کند و مولفه تولید پاسخ در مورد اینکه چگونه اطلاعات بیان شوند، تصمیم‌گیری می‌کند.

۳-۵-۶- انتشار خروجی

این مولفه، پیامی را که توسط بخش تولید پاسخ ساخته شده است را با توجه به نوع سیستم گفتگو، به فرمت مناسب ترجمه می‌کند. در سیستم‌های گفتگو متنی این مولفه وجود ندارد و پاسخ مستقیما از

[۱] Natural Language Query Processing system

واحد تولید پاسخ برای کاربر نمایش داده می‌شود. در سیستم‌های گفتگو صوتی برای انتشار خروجی دو روش استفاده می‌شود. اولین روش این است که از قالب‌های آماده برای ارسال پاسخ استفاده کرد. این قالب‌ها از طریق نمونه‌های بازیابی شده، ضبط شده‌اند. به طور مثال در ابتدای مکالمه قالب «خوش آمدید، چگونه می توانم به شما کمک کنم؟» برای مشتری نمایش داده می‌شود. رویکرد دوم استفاده از برنامه متن به گفتار[۱] است. در این روش، متن به صوت تبدیل می‌شود.

۳-۶- طبقه‌بندی سیستم‌های گفتگو

براساس نوع رهبری مکالمه، یک سیستم گفتمان می تواند به سه دسته تقسیم شود. منبع [۵۴] معتقد است معماری مدیر گفتگو توسط این سه دسته ساخته می‌شود.

- سیستم‌های گفتگو مبتنی بر حالت محدود (یا گراف)[۲]
- سیستم‌های گفتگو مبتنی بر قالب[۳]
- سیستم‌های گفتگو مبتنی بر عامل[۴]

۳-۶-۱- سیستم‌های گفتگو مبتنی بر حالت محدود

جریان گفتگو در این سیستم‌ها به عنوان یک مجموعه‌ای از حالت‌های گفتگو تعیین می‌شود. سیستم اطلاعات مورد نیاز را از کاربران در طی مراحل از پیش تعیین شده می‌گیرد. شکل (۲-۳) مثالی از نحوه عملکرد این سیستم را نمایش می‌دهد.

پیاده‌سازی این سیستم آسان است و واژگان و دستور زبان مورد نیاز برای هر حالت را می‌توان از پیش تعیین کرد. اما این نوع سیستم‌ها معایب زیادی دارند. مکالمه‌ها در این نوع سیستم‌ها حالت طبیعی ندارند. این سیستم‌ها به کاربر اجازه نمی‌دهند اطلاعات بیشتری را ارائه دهند و توانایی کاربر برای پرسیدن سوال و بدست گرفتن کنترل مکالمه را محدود می‌کنند.

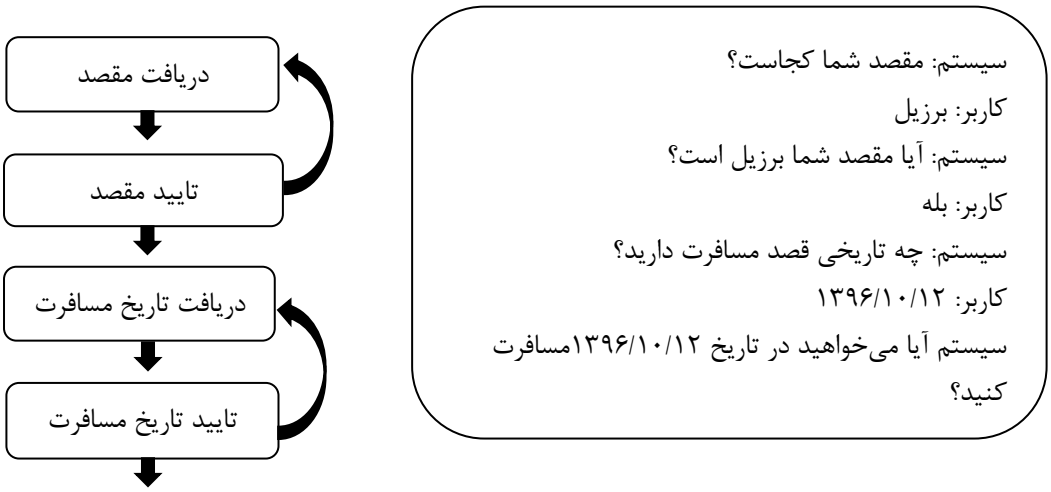
---

[۱] Text To Speech (TTS)

[۲] Finite-State DM

[۳] Frame/Form Based DM

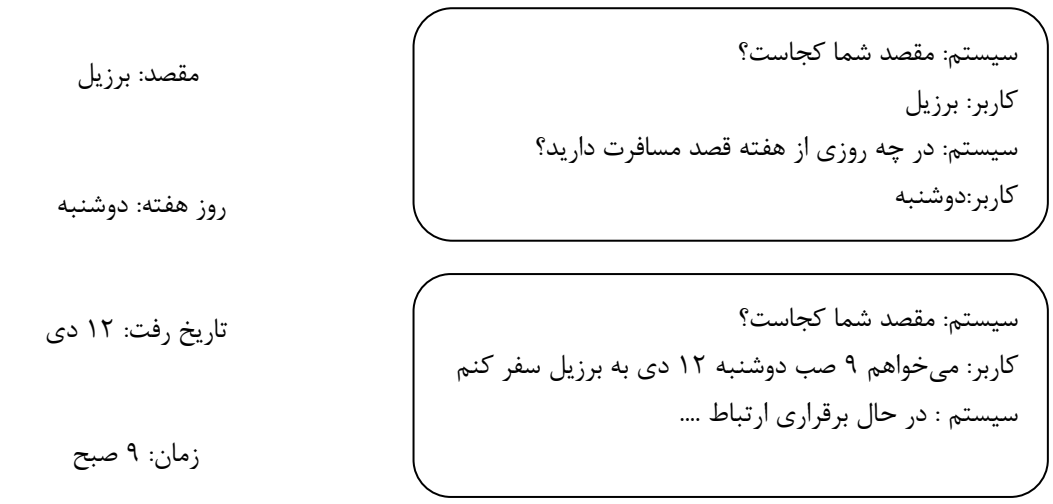
[۴] Agent Based Systems



شکل (۳-۲): مثالی از سیستم گفتگو مبتنی بر حالت محدود

۳-۶-۲- سیستم مبتنی بر قالب

سیستم‌های مبتنی بر قالب، فیلدهایی دارد که با استفاده از پاسخ کاربر پر می‌شود. جریان گفتگو در این نوع سیستم از پیش تعیین نشده است، اما بستگی به محتوای ورودی کاربر و اطلاعات دارد. شکل (۳-۳) مثالی از این نوع سیستم را نشان می‌دهد. در این شکل دو گفتگوی متفاوت وجود دارد در گفتگوی اول، مکالمه مشابه سیستم حالت محدود پیش می‌رود. در گفتگو دوم، کاربر در پاسخ به یک سوال اطلاعات بیشتری را ارائه می‌دهد، سیستم فیلدهایش را از ورودی کاربر پر می‌کند و اطلاعات باقی مانده را می‌پرسد.



شکل (۳-۳): مثالی از سیستم گفتگو مبتنی بر قالب

گفتگوها در این سیستم نسبت به سیستم حالت محدود، طبیعی‌تر هستند و کاربر می‌تواند پاسخ‌های آموزنده‌تری ارائه دهد. اما این سیستم‌ها نمی‌توانند گفتگوهای پیچیده را مدیریت کند و محدوده کاربرد، محدود به سیستم‌هایی است که اطلاعات را از کاربران استخراج می‌کنند.

۳-۶-۳- سیستم عامل مبتنی بر عامل

این سیستم‌ها اجازه می‌دهند سیستم، کاربر و برنامه با یکدیگر ارتباط پیچیده داشته باشند. تعامل در این سیستم به صورت ارتباط بین دو عامل دیده می‌شود که هر کدام قادر به استدلال در مورد اقدامات و باورهای خود هستند. مدل گفتگو با توجه به زمینه پیشین ساخته می‌شود. گفتگو به صورت پویا به عنوان دنباله ای از نوبت‌های مرتبط تکامل می‌یابد. این سیستم‌ها می‌توانند محدود به یک دامنه خاص نباشند. گفتگوها نسبت به دو سیستم قبل طبیعی‌تر هستند. اما ساختن این سیستم مشکل است زیرا پیاده‌سازی عامل پیچیده است.

۳-۷- سیستم‌های پرسش و پاسخ

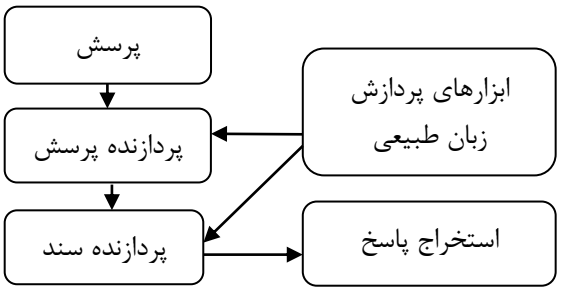
سیستم‌های پرسش و پاسخ به سیستم‌هایی گفته می‌شوند که به صورت خودکار و دقیق بتوانند سوال‌های پرسیده شده در حوزه زبان طبیعی را پاسخ دهند. این سیستم‌ها در زمینه‌هایی مانند بازیابی اطلاعات، پردازش زبان طبیعی و یادگیری ماشین قرار می‌گیرند. تاکنون بیشتر تحقیقات، برروی نوع سوال مطرح شده تمرکز کرده‌اند و سیستم‌های پرسش و پاسخ با توجه به نوع سوال دسته-بندی شده‌اند. یکی از مشهورترین تحقیقات، سیستم پرسش و پاسخ حقیقی[۱] است که در آن هدف پاسخگویی به پرسش‌های حقیقی است. در این نوع سیستم، کاربر پرسشی را مطرح می‌کند که پاسخ آن یک موجودیت ساده است مانند یک فرد، سازمان یا مکان. به اینگونه پرسش‌ها، حقیقی گفته می-شود. پرسش‌های دیگر از نوع، پرسش های تعریف [۵۵]، پرسش‌های نیازمند دلیل [۵۶]، پرسش‌های پیچیده [۵۷] و پرسش‌های تحلیلی [۵۸] هستند. پرسش‌های پیچیده به پرسش‌هایی گفته می‌شود که پاسخ آنها یکسان نیست و نیازمند ترکیب چندین پاسخ است. به عنوان مثال«پیامدهای جنگ

[۱] Factoid

تحمیلی چه بود؟». پرسش‌های تحلیلی، نیازمند انجام محاسبات و آنالیزهایی برای پاسخ دارند. به عنوان مثال «سن ملکه انگلستان چقدر است؟» برای پاسخ‌گویی بایستی تاریخ امروز از تاریخ تولد ملکه انگلستان کم شود تا سن واقعی او بدست آید.

۳-۸- معماری سیستم‌های پرسش و پاسخ

سیستم پرسش و پاسخ به طور کلی از سه مولفه پردازنده سوال، پردازنده سند و استخراج پاسخ تشکیل می‌شود که در شکل (۳-۴) نشان داده شده است. اکثریت سیستم‌های پرسش و پاسخ موجود، یک خط لوله‌ی یک طرفه هستند که هر کدام از ماژول ها به ترتیب اجرا می‌شوند و کاربر نمی‌تواند در همه‌ی موارد دخالت کند.



شکل (۳-۴): مولفه‌های اصلی سیستم پرسش و پاسخ

۳-۸-۱- پردازنده سوال

نقش پردازنده سؤال این است که از تکنیک‌های مختلف پردازش زبان طبیعی برای تفسیر سوال، استفاده کند. تفسیر سوال شامل تعیین دسته معنایی آن (مثلا شخص، سازمان، محل، و غیره) و استخراج مجموعه کلمات کلیدی، به منظور ایجاد لیستی از عبارت‌ها برای مؤلفه پردازنده سند است. پردازنده سوال سعی می‌کند تفسیر دقیق‌تری را با اضافه کردن اطلاعات معنایی به عبارت‌ها و شناسایی روابط هماهنگ بین آنها ایجاد کند. تا حدودی، پردازنده سؤال، مشابه مؤلفه فهم زبان طبیعی در یک سیستم گفتگو عمل می‌کند.

۳-۸-۲- پردازنده سند

پردازنده سند مجموعه‌ای از اسناد را استخراج می‌کند که بوسیله آن، پاسخ از طریق اطلاعات بازیابی شده از پاراگراف‌ها بدست می‌آید. با استفاده از خروجی پردازنده سؤال، پردازنده سند پاراگراف‌هایی را که می تواند به سوال کاربر پاسخ دهد را پیدا می‌کند. اغلب پردازش‌های انجام شده در این مرحله بسیار کم است و تنها محدودیت‌هایی بر روی پاراگراف‌های بازیابی شده اعمال می‌کند. پاراگراف‌های بازیابی شده بایستی شامل بیشتر کلمات کلیدی و حداقل یک کلمه در رده مشابه پاسخ مورد انتظار باشند. با این حال، سیستم‌هایی وجود دارند که لیستی از عبارت‌هایی که توسط پردازنده سوال با اصطلاحات معناشناختی بازگردانده شده است را گسترش می‌دهند و معیارهای پیشرفته‌ای را که ضریب اطمینان پاراگراف حاوی پاسخ را اندازه‌گیری می‌کند را پیاده‌سازی می‌کنند.

۳-۸-۳- استخراج پاسخ

ماژول استخراج پاسخ، پاراگراف‌هایی را که توسط پردازنده سند، بازیابی شده است را دریافت می‌کند و پاسخ واقعی پرسش را پیدا می‌کند. این کار معمولا با استفاده از دو نوع الگوریتم استخراج الگو پاسخ و n-گرم‌ها به دست می‌آید. روش استخراج الگو پاسخ[۱]، متکی بر الگوهای دستی یا خودکار برای شناسایی پاسخ یک سوال است. این الگوها می‌توانند سطحی یا براساس اطلاعات نحوی و معنایی استخراج شده توسط پردازنده سؤال باشند. روش n-گرم‌ها از یک مجموعه بزرگ از اسناد مانند وب برای استخراج قطعه‌های تکراری استفاده می‌کند. سپس این قطعه‌ها با توجه به احتمال پاسخ بودن سوال امتیازدهی می‌شوند. برای سیستم‌های پرسش و پاسخ که فراتر از پرسش و پاسخ حقیقی هستند، مرحله استخراج پاسخ می‌تواند اطلاعات ترکیبی از چندین اسناد را برای تولید پاسخ فراهم کند. از این منظر، این مرحله شبیه به مولفه‌های تولید پاسخ در سیستم‌های گفتگو است. در برخی موارد، از یک ماژول اعتبارسنجی پاسخ، برای بررسی ارزش پاسخ بازیابی شده، استفاده می‌شود.

[۱] Answer-Type Pattern Extraction

۳-۹- سیستم پرسش و پاسخ تعاملی

در پرسش و پاسخ تعاملی، کاربران یک گفتگوی گسترده‌ای با سیستم تشکیل می‌دهند تا اطلاعات مربوط به یک سناریوی پیچیده را بدست آورند. با توجه به اینکه اطلاعات مورد نظر با شیوه‌ای تکراری از طریق پرسیدن سؤالات مختلف ساده بدست می‌آیند، سیستم‌های پرسش و پاسخ تعاملی دقیق‌تر از سیستم‌های پرسش و پاسخ معمول هستند. دلیل این امر این واقعیت است که سیستم‌های پرسش و پاسخ تعاملی محدودیت‌ها را در سوال‌ها مشخص می‌کنند و با کاربران تا زمانی که این محدودیت‌ها اصلاح شوند، تعامل می‌کند. به بیانی دیگر، سیستم‌های پرسش و پاسخ تعاملی به احتمال کمی باید با ساختارهای مبهم و پیچیده مقابله کنند. همچنین زمانی که یک سیستم پرسش و پاسخ تعاملی با چنین ساختار مبهمی مواجه شد، می‌تواند یک گفتگو را برای روشن شدن درخواست کاربر آغاز کند. هنگامی که سیستم، پاسخ‌های بیش از حد زیاد و یا کمی برای یک سوال دارد ازکاربر درخواست می-شود که سوال خود را بازنویسی کند. سیستم پرسش و پاسخ تعاملی به کاربران امکان می‌دهد سوالات خود را مطرح کنند و بر اساس بازخورد دریافت شده از سیستم، سوال خود را اصلاح کنند؛ تا زمانی که پاسخ مورد نیاز را پیدا کنند. از این دیدگاه، پرسش و پاسخ تعاملی باید به عنوان یک مکالمه با رهبری دو طرفه در نظر گرفته شوند. به این دلیل که حتی اگر کاربر کسی است که به طور معمول سوال‌ها را مطرح می‌کند، ممکن است سیستم زمانی که نیاز به توضیح دارد، کنترل را بدست بگیرد. تعامل با سیستم‌های پرسش و پاسخ تعاملی همانند سیستم‌های گفتگو از طریق مکالمه انجام می‌شود. بنابراین هر سیستم پرسش و پاسخ تعاملی می تواند یک سیستم گفتگو در نظر گرفته شود. اما تنها سیستم‌های گفتگو که مبتنی بر جستجوی اطلاعات هستند یا به عبارتی دیگر، هدفمند هستند، به عنوان سیستم های پرسش و پاسخ تعاملی در نظر گرفته می‌شوند. تعامل بین سیستم‌های پرسش و پاسخ تعاملی و کاربران انسانی بیشتر مبتنی بر وظیفه است و معمولا شامل نوبت‌های کمتر می‌شود. همچنین بر خلاف یک سیستم گفتگو، معمولا از خصوصیت‌های مکالمه انسانی برخوردار نیست، بنابراین تعامل با آنها شکلی طبیعی ندارد. هدف از تعامل با سیستم‌های پرسش و پاسخ تعاملی یافتن

اطلاعات است، به همین دلیل مکالمه‌های این سیستم‌ها ساختار واضحتری نسبت به سیستم‌های گفتگو دارند و در بیشتر موارد واضح است که نوبت بعدی کیست. در هر دو سیستم گفتگو هدفمند و سیستم‌های پرسش و پاسخ، زمینه را می‌توان با محدود کردن دامنه، ایجاد کرد. به این ترتیب، هر دو شرکت کننده انسانی و رایانه می‌توانند گفته‌های مربوط به آن حوزه را تفسیر کنند و میزان ابهام در پرسش‌ها و پاسخ‌ها را کاهش دهند.

۳-۹-۱-انواع سیستم‌های پرسش و پاسخ تعاملی

زمینه پرسش و پاسخ تعاملی کاملا جدید است و سیستم‌های زیادی وجود ندارد. به همین دلیل، سیستم‌های موجود از نظر ساختار و راه حل مورد استفاده بسیار متفاوت هستند. دو راه برای توسعه این سیستم‌ها وجود دارد. نسخه اول با تولید یک نسخه سطح پایین از یک سیستم گفتگوی مبتنی بر جستجوی اطلاعات بدست می‌آید. نسخه دوم با ادغام ویژگی‌های اضافی به یک سیستم پرسش و پاسخ استاندارد، ایجاد می‌شود.

۳-۹-۲-اعلان مکالمه

تجزیه و تحلیل ساختار مکالمه در قالب اعلان مکالمه انجام می‌شود. اعلان مکالمه، اطلاعات مهمی در مورد جریان و محتوای مکالمه فراهم می‌کند و می‌تواند به عنوان اولین قدم برای توسعه عامل‌های هوشمند در سیستم‌های با کنترل دوطرفه، استفاده شود. داشتن یک درک غنی از ساختار مکالمه فراتر از سؤال‌ها و پاسخ‌ها می‌تواند استخراج پرسش و پاسخ را بهبود بخشد. پیش‌بینی اعلان مکالمه می‌تواند برای هدایت خودکار تولید پاسخ استفاده شود. توسعه مدل‌های اعلان مکالمه یکی از زمینه-های مشهور در جهت بررسی ساختار مکالمه است. منبع [۵۹] از دو مجموعه اعلان مکالمه درشت‌دانه و ریزدانه استفاده می‌کند. در دسته درشت‌دانه، ۱۰ اعلان مکالمه وجود دارند که عبارتند از: باز و بسته کردن مکالمه هنگام خوش‌آمدگویی، اظهارات مثبت و منفی، انواع مختلف سوالات، انواع پاسخ‌های مثبت و منفی، واکنشی و قدردانی. مجموعه اعلان‌های مکالمه ریزدانه به دو دسته گسترده بیانیه‌ها و درخواست‌ها تقسیم می‌شود. منبع [۴۶] چهار دامنه مختلف از مکالمه‌ها را بررسی کرده است. دامنه-

های آزمایش شده عبارتند از: مکالمات ایمیلی با ۱۲ اعلان مکالمه، انجمن با ۱۱ اعلان مکالمه، جلسه با ۱۱ اعلان مکالمه و مکالمات تلفنی با ۱۶ اعلان مکالمه.

۳-۹-۳- ارتباط بین نوبت‌ها در مکالمه

قطعه‌های بامعنی نوبت‌ها با اعلان‌های مکالمه، همبستگی قوی دارند. به عنوان مثال، اگر گوینده شماره ۱، یک سوال برای گوینده شماره ۲ مطرح کند، احتمال بالایی دارد که گفتگو بعدی مکالمه، پاسخی از گوینده شماره ۲ باشد. بنابراین ساختار مکالمه عامل مهمی است که باید برای مدل‌سازی اتوماتیک اعلان‌های مکالمه در نظر گرفته شود. منبع [۴۱] ساختار مکالمه را توسط روابط بین اعلان-های مکالمه تعریف می‌کند. اعلان‌هایی که در نظر می‌گیرد، پرسش، پاسخ، اطلاعیه، توافق، عدم توافق، قدردانی، واکنش منفی، توضیحات و طنز هستند. برای پاسخ، توافق، عدم توافق، قدردانی، واکنش منفی، توضیحات و طنز یک نوبت قبل را به عنوان سابقه گفتگو در نظر می‌گیرد. منبع [۱۳] میزان تاثیر نوبت‌های پیشین را با نام حافظه بررسی می‌کند. این منبع اندازه حافظه را ثابت در نظر می‌گیرد.

۳-۹-۴-ارزیابی

همانند سایر روش‌های پردازش متن، ارزیابی نقش مهمی را در سیستم‌های پرسش و پاسخ تعاملی ایفا می‌کند. با این حال، هیچ روش خاصی در این زمینه وجود ندارد و تنها گزینه‌های موجود استفاده از روش‌های ارزیابی در سیستم‌های پرسش و پاسخ استاندارد و سیستم‌های گفتگو است. بر خلاف سیستم‌های پرسش و پاسخ معمول، که عملکرد یک سیستم براساس قدرت پاسخ‌گویی به یک سوال ارزیابی می‌شد، عملکرد سیستم‌های پرسش و پاسخ تعاملی توسط تجزیه و تحلیل مشخصه‌های گفتگو بررسی می‌شود. مشکل اصلی در طراحی یک روش ارزیابی، این است که به ندرت پیش‌بینی می‌شود که چگونه مکالمه تعاملی به صورت موفق تمام می‌شود. به همین دلیل ضروری است که افراد در فرایند ارزیابی دخیل باشند که باعث می‌شود که روند کند، گران و دشوار باشد.

ارزیابی سیستم‌های پرسش و پاسخ استاندارد بسته به اینکه آنها چه نوع سوالی را ارزیابی می‌کنند، متفاوت است. یک راه حل، ایجاد مجموعه‌ای از پرسش و پاسخ‌های استاندارد است که بوسیله آن بتوان سیستم را ارزیابی کرد. این روش برای سؤال‌های پیچیده قدرتمند نیست. زیرا پاسخ اینگونه پرسش‌ها می‌تواند از طریق روش‌های مختلف بیان شود یا حتی می‌تواند چندین جواب ممکن باشد. در این مورد، ارزیابی‌های انسانی (مانند مسابقات ارزیابی TREC) مورد استفاده قرار می‌گیرند. تکنیک-های ذکر شده همچنین می‌توانند برای پاسخ دادن به سیستم‌های تعاملی برای بررسی صحت پاسخ-های برگردانده شده به کاربران استفاده شوند، اما به تنهایی اطلاعات کافی در مورد کیفیت سیستم را ارائه نمی‌دهند. به همین دلیل، روش‌های ارزیابی سیستم‌های گفتگو نیز مورد استفاده قرار می‌گیرد.

ارزیابی سیستم‌های گفتگو نیز یک وظیفه مشکل است. بعضی منابع مانند [۵۱] از ارزیابی مبتنی بر وظیفه استفاده می‌کنند که در آن کیفیت یک سیستم گفتگو براساس تکمیل وظیفه سنجیده می‌شود. منبع [۵۳] ارزیابی را در سه قسمت انجام می‌دهد:

- **موفقیت مکالمه:** راه حل صحیح توسط سیستم گفتگو یافت می‌شود؛
- **هزینه بهره‌وری:** میزان زمان صرف شده برای یافتن یک راه حل. به عنوان مثال تعداد نوبت؛
- **هزینه‌های کیفیت:** تا چه میزان سیستم خوب کار می‌کند، مثلا تعداد دفعاتی که کاربر مجبور است سیستم را قطع کند.

با توجه به منبع [۶۰]، در سیستم پرسش و پاسخ تعاملی، مکالمه‌ها معمولا با توجه به موارد زیر ارزیابی می‌شوند:

- **کارایی:** تعداد سوالاتی که کاربر باید برای یافتن اطلاعات خاص، بپرسد.
- **عملکرد:** مرتبط بودن پاسخ‌های بازگشتی؛
- **رضایت کاربر:** بازخوردهای که کاربر به سیستم نشان می‌دهد.

با توجه به این که از پرسشنامه‌ها برای ارزیابی سیستم‌های گفتگو استفاده می‌شود، طبیعی است که این استراتژی بر روی سیستم‌های پرسش و پاسخ تعاملی نیز منطبق می‌شود. این کار، فرصتی را برای ارزیابی جنبه‌های مختلف سیستم فراهم می‌کند.

منبع [۱۸] از حاشیه نویسان انسانی می‌خواهند سوال‌های جایگزین سیستم را در مواردی که سیستم نمی‌تواند به یک سوال پاسخ دهد، رتبه‌بندی کنند. به این ترتیب، آنها سعی می‌کنند دریابند که کدام محدودیت‌ها را کاربر به منظور بدست آوردن یک پاسخ برای اولین بار تغییر می‌دهد.

۳-۱۰- جمع‌بندی

در این فصل به بررسی مکالمه‌ها و سیستم‌هایی که از طریق تعامل با کاربر، نیاز اطلاعاتی آن‌ها را برطرف می‌کردند، پرداخته شد. در ابتدا ویژگی‌های مکالمه‌های انسانی بررسی شدند. این ویژگی‌ها از آن جهت مفید هستند که به ایجاد یک مکالمه طبیعی با انسان کمک می‌کنند. پس از آن سیستم‌های گفتگو معرفی شدند. این نوع سیستم‌ها به دو نوع هدفمند و بدون هدف تقسیم می‌شوند. معماری کلی این نوع سیستم‌ها در ادامه توضیح داده شد. پس از آن، به توضیح سیستم‌های پرسش و پاسخ پرداخته شد و معماری آن‌ها شرح داده شد. سیستم‌های پرسش و پاسخ تعاملی از تقاطع بین سیستم-های گفتگو و سیستم‌های پرسش و پاسخ، بوجود آمدند. در پایان این نوع سیستم‌ها و نحوه عملکرد آن‌ها به طور کامل تحلیل شدند و روش‌های ارزیابی این سیستم‌ها معرفی شدند.

فصل چهارم:

استفاده از گفتگوی پیشین در دسته‌بندی پاسخ‌ها

(روش پیشنهادی)

۴-۱- مقدمه

دسته‌بندی قطبی یکی از زمینه‌های مهم در پردازش متن است. با استفاده از این زمینه می‌توان تفکر درونی یک متن را دریافت کرد. دسته‌بندی‌های مختلفی که تاکنون استفاده شده‌اند، دسته متن را بدون سابقه آن مشخص می‌کنند. همچنین تاکنون این زمینه بر روی متن‌های تعاملی مانند مکالمه‌ها، انجام نشده است. در این فصل به معرفی روش‌های پیشنهادی برای دسته‌بندی پاسخ‌های یک سیستم پرسش و پاسخ تعاملی پرداخته می‌شود. این تحقیق نقطه شروعی برای بررسی تاثیر گفته‌های پیشین در دسته‌بندی قطبی است. روش پیشنهادی این پایان‌نامه شامل چند مرحله است. ابتدا با استفاده از روش‌های پیش‌پردازش متداول در پردازش متن، داده‌های خام پیش‌پردازش می‌شوند و نتیجه پیش-پردازش‌های مختلف با یکدیگر مقایسه می‌شوند. در مرحله بعدی مجموعه ویژگی‌ها و توابع وزن‌دهی مربوط به آن‌ها بررسی خواهند شد. برای بررسی میزان موثر بودن مجموعه ویژگی‌ها، از روش‌های یادگیری ماشین با نظارت استفاده می‌شود. سپس از روش‌های مشهور کاهش ویژگی و بدست آوردن عبارت‌های مفید در دسته‌بندی استفاده می‌شود. تمام این مراحل مشابه دسته‌بندی قطبی متداول در متن‌هاست. مهمترین قسمت، بررسی سابقه مکالمه در دسته‌بندی پاسخ است که به دو صورت وابستگی زمانی و وابستگی ساختاری تحلیل می‌شود. در نهایت روش‌های مشهور ارزیابی در پردازش متن معرفی می‌شوند.

شکل (۴-۱): فرایند دسته‌بندی پاسخ‌ها

۴-۳- پیش‌پردازش

مکالمه‌هایی که توسط کاربر و سیستم مبادله می‌شوند، همانند هر داده خام دیگر، حاوی ناخالصی‌های زیادی می‌باشند. این ناخالصی‌ها به صورت غلط‌های املائی، حالت‌های اختصار کلمه، حالت‌های نامتعادل و غیره به متن‌ها اضافه می‌شوند [۲۰]. در ابتدا باید پردازش‌های اولیه‌ای بر روی منابع خام انجام داد تا بتوان ناخالصی‌ها و نویزهایی که دقت را کاهش‌می‌دهند را حذف کرد.

علائم نگارشی: وجود علائم نگارشی مانند نقطه، علامت سوال و تعجب می‌توانند نوع جمله را مشخص کنند. این نوع علائم نگارشی برای آنالیز احساسی در سطح جمله بسیار مفید می‌باشند. زیرا بوسیله آنها می‌توان جمله‌های یک متن را تشخیص داد و براساس آنها دسته‌بندی را انجام داد. در بعضی موارد کاربران در استفاده از این علائم، اشتباه می‌کنند. به عنوان مثال در هنگام پرسش فراموش می‌کنند که علامت سوال را در انتهای جمله اضافه کنند. در حالت‌هایی دیگر درج علامت سوال به معنای سوالی بودن پیام نیست. بدلیل چنین مشکلاتی بهتر است این علامت‌ها پیش از پردازش، حذف شوند.

واژگان اضافی[۱]: در هر متنی کلماتی وجود دارد که به تعداد زیاد در متن تکرار می‌شوند و با حذف آن‌ها معنی متن تغییر نمی‌کند. به این نوع کلمه‌ها، واژگان یا کلمه‌های اضافی گفته می‌شود. به عنوان مثال در زبان فارسی با، در، از، یا، و، به و در زبان انگلیسی on، of، and و or کلمه‌های اضافی هستند. اکنون برای زبان‌های مختلف، مجموعه کلمه‌های اضافه به صورت آماده موجود است. مهم‌ترین مشکل این مجموعه آماده این است که حاوی کلمات نفی هستند. کلمات نفی در دسته‌بندی قطبی بسیار مهم هستند. منبع [۲۰] از شش روش متداول بازیابی اطلاعات برای حذف کلمات اضافی استفاده کرد و به این نتیجه رسید که استفاده از حذف کلمات اضافی به دلیل خلوت بودن متن‌ها، روشی نامناسب برای دسته‌بندی قطبی است. در این تحقیق برای حذف کلمات اضافی از کتابخانه استندفورد استفاده می‌شود.

[۱] Stop Words

ریشه‌یابی کلمه‌ها: یک کلمه می‌تواند به شکل‌های مختلفی در متن نمایش داده شود. این شکل‌های مختلف با استفاده از پسوندها و پیشوندهای مختلف بوجود می‌آیند. برای افزایش دقت، بهتر است که کلیه کلمه‌ها به حالت پایه‌ای خود برگردند. ابزارهای متنوعی برای ریشه‌یابی وجود دارد. در این پایان-نامه از ریشه‌یاب وکا[1] استفاده شده است.

استفاده از شکل‌های یکسان حروف: در بعضی از زبان‌ها برای یک حرف، حالت‌های مختلفی وجود دارد به عنوان مثال در زبان انگلیسی هر حرف به دو شکل کوچک و بزرگ نمایش داده می‌شود. با در نظر گرفتن تنها یک شکل از حرف می‌توان دقت را بالا برد و از تشخیص اشتباه جلوگیری کرد.

حذف حالت‌های نامتداول: در بعضی اوقات، کاربران برای تاکید بیشتر برروی یک کلمه، بعضی از حرف‌های آن را تکرار می‌کنند. برای استفاده از این‌گونه کلمه‌ها بایستی این حروف اضافی پیدا شوند و حذف گردند.

۴-۴- ویژگی‌های انتخابی

فرکانس عبارت، حضور عبارت، TF-IDF کلاسیک، ویژگی احتمالی و نسخه تغییریافته اطلاعات متقابل نقطه به نقطه، ویژگی‌های انتخابی هستند. این ویژگی‌ها بر روی n-gram ها آزمایش شدند. با استفاده از bigram می‌توان تاثیر ترکیب کلمات نزدیک بهم را بررسی کرد. زیرا بسیاری از کلمه‌ها در کنار یکدیگر، قطبی متفاوت در پاسخ ایجاد می‌کنند. به عنوان مثال «no problem» پاسخی مثبت است هرچند که از دو کلمه منفی «no» و «problem» ایجاد شده است. در این پایان نامه فرض می-شود استفاده از bigram موجب بهبود نتایج روش‌های دسته‌بندی شود. سه ویژگی فرکانس عبارت، حضور عبارت و TF-IDF به طور کامل در فصل دوم بررسی شدند. TF-IDF کلاسیک از ضرب فرکانس عبارت و فرکانس معکوس عبارت کلاسیک بدست می‌آید. رابطه (۴-۱)، TF-IDF کلاسیک را نشان می‌دهد.

[1] Weka Stemmer

$$TF.IDF = tf \times \log \frac{N}{df} \qquad (۱-۴)$$

همانطور که در فصل اول توضیح داده شد، پاسخ‌ها به صورت‌های مختلفی بیان می‌شوند. بنابراین کلمه‌ها ممکن است در هر سه دسته مشترک باشند. با این حال بعضی از کلمه‌ها در یک دسته بیشتر تکرار می‌شوند. به این کلمه‌ها که مختص دسته خاصی هستند، کلمه‌های احساسی می‌گویند. این کلمه‌ها به صورت آماده برای بعضی از زبان‌ها وجود دارد. محققان از این کلمه‌ها برای شناسایی جهت معنایی کلمات ناشناخته استفاده می‌کنند. با این حال این گونه کلمات برای بسیاری از زبان‌ها، موجود نیست. استفاده از ویژگی که مشخص کند کلمه‌ها در چه دسته‌هایی بیشتر تکرار می‌شوند، در دسته-بندی قطبی بسیار ارزشمند است. بایستی با استفاده از روش‌های آماری بتوان ویژگی را ایجاد کرد که به کلمه‌های احساسی ارزش بیشتری بدهند. ویژگی احتمالی و نسخه تغییریافته اطلاعات متقابل نقطه به نقطه، ویژگی‌های پیشنهادی هستند.

۴-۴-۱- احتمال رخداد یک عبارت در یک دسته (ویژگی احتمالی)

ویژگی احتمالی، یک عبارت را براساس بیشترین تکرار در متن‌های یک دسته، مقداردهی می‌کند. به عبارت دیگر گرایش بیشتر عبارت را به یکی از دسته‌ها در نظر می‌گیرد و اهمیت عبارت را براساس آن محاسبه می‌کند. این ویژگی از طریق رابطه (۴-۲) محاسبه می‌شود.

$$prob(t) = \frac{\max(tf_{pos}(t), tf_{neg}(t), tf_{neu}(t))}{tf_{pos}(t) + tf_{neg}(t) + tf_{neu}(t)} \qquad (۲-۴)$$

در رابطه (۴-۱)، t عبارت است؛ $tf_{pos}(t)$، تعداد تکرار عبارت t در دسته متن‌های مثبت، $tf_{neg}(t)$، تعداد تکرار عبارت t در دسته متن‌های منفی و $tf_{neu}(t)$، تعداد تکرار عبارت t در دسته متن‌های خنثی را نشان می‌دهد. مثال زیر نحوه محاسبه این ویژگی را نشان می‌دهد.

به طور کلی کلمه «not» در زبان انگلیسی یک کلمه احساسی منفی است. ولی این کلمه در متن‌های دسته مثبت و دسته خنثی نیز تکرار می‌شود. به عنوان مثال در متن مثبت «not bad» و متن خنثی «I do not know». اگر این کلمه ۵ بار در متن‌های آموزشی مثبت، ۱۵ بار در متن‌های خنثی و ۳۰

بار در متن‌های منفی تکرار شده باشد، ارزش کلمه «not» برابر ۰/۶ می‌شود. زیرا بیشترین تعداد تکرار را در کلاس منفی داشته که با تقسیم آن بر روی تعداد کل تکرارها که مقدار ۰/۶ بدست می‌آید.

۴-۴-۲- اطلاعات متقابل نقطه به نقطه

همانطور که در فصل دوم توضیح داده شد از این تابع بیشتر به عنوان روشی برای تعیین جهت معنایی یک کلمه ناشناخته، استفاده می‌شود [۳, ۲۶]. با اندکی تغییرات می‌توان از این تابع به عنوان یک ویژگی قدرتمند در n-گرم‌ها استفاده کرد. رابطه (۴-۳)، تابع توسعه یافته اطلاعات متقابل نقطه به نقطه را معرفی می‌کند:

$$PMI(t) = \log\left(\max \frac{P(t,c_i)}{P(t)\times P(c_i)}\right) \quad i = pos, neg, neu \qquad (۴-۳)$$

$P(t,c_i)$ احتمال رخداد عبارت t در متن‌های دسته c_i است؛ $P(c_i)$ احتمال رخداد دسته c_i را نشان می‌دهد و $P(t)$ احتمال رخداد عبارت t در کل متن‌ها است.

۴-۵- روش‌های دسته‌بندی

با توجه به کاربرد گسترده و عملکرد بالا روش‌های یادگیری ماشین بانظارت در دسته‌بندی قطبی، دسته‌بندی پاسخ‌ها از این طریق انجام می‌شود. برای دسته‌بندی داده‌ها از نرم‌افزار وکا استفاده شده است. روش‌هایی که در این پایان‌نامه استفاده شدند، عبارتند از بیزین ساده، ماشین بردار پشتیبان، درخت تصمیم و الگوریتم Bagging (درخت تصمیم).

۴-۵-۱- بیزین ساده

کلاس‌بندی بیزین ساده بدلیل اینکه از نظر هزینه محاسباتی به پردازنده و حافظه کمی نیاز دارد، بسیار کارآمد است. همچنین این روش نیاز به مقدار کمی از داده‌ها برای آموزش دارد و زمان آموزش نسبت به سایر روش‌ها کمتر است.

ایده اصلی بیزین ساده، پیداکردن احتمال دسته‌های متنی با استفاده از احتمال‌های اتصال ویژگی و دسته‌ها است. احتمال ویژگی X و دسته C برابر است با:

$$P(C \mid X) = \frac{P(X \mid C) \times P(C)}{P(X)} \qquad (۴-۴)$$

به بیانی دیگر بیزین ساده یک طبقه‌بندی تولیدی[۱] است که احتمال‌های پیشین $P(X \mid C)$ و $P(C)$ را از داده‌های آموزشی برآورد می‌کند و احتمال پسین $P(C \mid X)$ را بر اساس احتمالات پیشین تولید می‌کند.

۴-۵-۲-ماشین بردار پشتیبان

ماشین بردار پشتیبان، یک روش طبقه‌بندی جدا کننده[۲] است که هیچ پیش‌فرضی را بر اساس داده-های آموزشی ایجاد نمی‌کند و به طور مستقیم $P(C \mid X)$ را تخمین می‌زند. به طور کلی این روش برای جداسازی دو کلاس طراحی شده است. برای انطباق آن با کلاس‌های بیشتر از دو روش یک کلاس در برابر دیگران[۳] و یک کلاس در برابر یک کلاس[۴] استفاده می‌شود. روش یک کلاس در برابر دیگران، یک کلاس را به صورت تصادفی انتخاب می‌کند و باقی کلاس‌ها را به صورت یک کلاس در نظر می‌گیرد. سپس با استفاده از ماشین بردار پشتیبان کلاس‌بندی را انجام می‌دهد. در روش یک کلاس در برابر یک کلاس، ماشین بردار پشتیبان را بین دو کلاس از بین کلاس‌های مختلف اعمال می‌کند. این کار را آنقدر ادامه می‌دهد که کلیه کلاس‌ها از یکدیگر جدا شوند.

در این پایان‌نامه از سه تابع هسته خطی، چندجمله‌ای و تابع شعاعی پایه[۵] برای آزمایش ماشین بردار پشتیبان استفاده شده است. به دلیل اینکه این روش باید به صورت سه کلاسه باشد، از روش یک کلاس در برابر دیگران استفاده می‌شود. ماشین بردار پشتیبان با تابع هسته خطی در پردازش متن، بسیار کارآمد است. علت این امر این حقایق است که بیشتر مسائل دسته‌بندی متن به صورت خطی قابل جداسازی هستند. همچنین ابعاد مجموعه ویژگی‌هایی که در دسته‌بندی متن استفاده می‌شود، بالاست و نگاشت داده‌ها به یک فضایی با ابعاد بیشتر، کارایی را بهبود نمی‌دهد. سرعت آموزش تابع

---

[۱] Generative Classifier

[۲] Discriminative Classifier

[۳] One vs all

[۴] One vs one

[۵] Radial Basis Function(RBF)

هسته خطی نسبت به سایر توابع بیشتر است. همچنین پارامترهایی که باید در این تابع بهینه شوند، کمتر از باقی توابع است [۶۱].

۴-۵-۳- درخت تصمیم

درخت تصمیم روشی برای نمایش یک مجموعه از قوانین است که منتهی به یک دسته یا مقدار می‌شوند. درخت تصمیم از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شوند و هدف در این فرایند، افزایش فاصله بین گروه‌ها در هر جداسازی است. این روش ساختار ساده و قابل فهمی دارد و با ویژگی‌های پیش‌بینی کننده، بخوبی کار می‌کنند. اگر به درخت اجازه داده شود که بدون محدویت رشد کند، علاوه بر افزایش زمان ساخت، ممکن است داده‌ها بیش‌برازش[۱] شوند. برای جلوگیری از بیش برازش دو روش وجود دارد. در روش اول باید قبل از رسیدن به مرحله‌ای که داده‌های آموزشی به طور کامل دسته‌بندی شوند، رشد درخت را متوقف نمود. در روش دوم به درخت اجازه داده می‌شود که کامل رشد کند و سپس شاخه‌هایی که مفید نیستند هرس می‌شوند. در این پایان نامه از الگوریتم C4.5 استفاده شده است که توسط معیار اطمینان[۲] ۰/۲۵ شاخه‌ها را هرس می‌کند و حداقل تعداد هر نمونه در هر برگ برابر ۲ است.

۴-۵-۴- الگوریتم Bagging (درخت تصمیم)

با ترکیب نتایج تعدادی از الگوریتم‌های دسته‌بندی، می‌توان به دقت بالایی دست یافت. این الگوریتم‌های یادگیری، ممکن است در الگوریتم، پارامترها، نحوه نمایش(ویژگی‌هایی که استفاده می‌کنند) و مجموعه داده‌های آموزشی متفاوت باشند. روش‌های مختلفی برای ترکیب دسته‌بندی‌کننده‌ها وجود دارد. میانگین‌گیری و استفاده از رأی اکثریت نمونه‌ای از متداول‌ترین روش‌ها در ترکیب دسته‌بندی کننده‌ها است. الگوریتم Bagging مبتنی بر رأی‌گیری است با این تفاوت که یادگیری‌های پایه با داده‌های آموزشی متفاوتی، آموزش داده می‌شوند. این داده‌های آموزشی از طریق نمونه‌گیری تصادفی

[۱] Overfitting
[۲] Confidence Factor

با جایگزین از داده‌های آموزشی اصلی بدست می‌آیند. به هر مجموعه از این نمونه‌های آموزشی کیسه گفته می‌شود. روش Bagging برای دسته‌بندی‌کننده‌های ناپایدار یعنی الگوریتم‌هایی که با تغییر داده، دچار تغییر در نتیجه می‌شوند (مانند درخت تصمیم) عملکرد خوبی خواهد داشت. در این پایان‌نامه از درخت تصمیم برای دسته‌بندی هر کیسه استفاده می‌شود و تعداد کیسه‌ها ۱۰ عدد در نظر گرفته می‌شود. علت استفاده از درخت تصمیم در الگوریتم Bagging این است که هر درخت جداگانه در این الگوریتم مشکل بیش‌برازش نخواهند داشت.

۴-۶- روش‌های کاهش ویژگی

مهم‌ترین مزایای استفاده از الگوریتم‌های کاهش ویژگی این است که آموزش را سریع‌تر می‌کنند. همچنین دقت را با حذف ویژگی‌های نویزی افزایش می‌دهند و از بیش‌برازش جلوگیری می‌کنند. در این پایان‌نامه از روش‌های کاهش ویژگی بهره اطلاعات، نرخ بهره و مربع کای استفاده می‌شود. این روش‌ها، هر ویژگی را با استفاده از منطق آماری خود، رتبه‌بندی می‌کنند و همانطور که در فصل دوم توضیح داده شد، نیازمند یک روش جستجو برای پیدا کردن تعداد مناسب ویژگی‌ها دارند. روش جستجویی که برای هر سه روش استفاده شده است، روش رتبه‌بندی[۱] است که ویژگی‌ها را براساس ارزیابی فردی آن‌ها توسط روش‌های استخراج ویژگی ذکر شده، رتبه‌بندی و مرتب می‌کند. همچنین این روش با استفاده از یک حد آستانه بعضی از ویژگی‌های نامناسب را دور می‌ریزد. حالت پیش‌فرض آن به این صورت است که از آستانه استفاده نمی‌کند و تمامی ویژگی‌ها را رتبه‌بندی می‌کند. در این پایان‌نامه نیز از روش رتبه‌بندی پیش‌فرض برای هدف رتبه‌بندی ویژگی‌ها، استفاده می‌شود.

۴-۶-۱ بهره اطلاعات

بهره اطلاعات یک ویژگی، عبارت است از مقدار کاهش آنتروپی که بوسیله جداسازی نمونه‌های آموزشی، از طریق این ویژگی بدست می‌آید. به عبارت دیگر بهره اطلاعات یک مقدار آماری است که

[۱] Ranker

نشان می‌دهد یک ویژگی تا چه میزان قادر است نمونه‌های آموزشی را بر اساس نوع دسته آن‌ها جداکند.

۴-۶-۲- نرخ بهره

اگر یک ویژگی مقادیر گسترده‌ای داشته باشد، دارای بهره اطلاعات بالایی خواهد بود. این بهره اطلاعات بالا، موجب می‌شود که این ویژگی به عنوان ریشه درخت انتخاب شود و درخت حاصل عمق کمی داشته باشد. در نتیجه این ویژگی به تنهایی می‌تواند تمام نمونه‌های آموزشی را دسته‌بندی کند. با این وجود، این ویژگی در مقابل داده‌های دیده نشده یا آزمون، به خوبی داده‌های آموزش دیده، عمل نمی‌کند. برای جلوگیری از رتبه‌دهی بالا به ویژگی‌هایی که مقادیر گسترده و یکنواخت دارند از نرخ بهره استفاده می‌شود. نرخ بهره یک نسخه توسعه یافته از بهره اطلاعات است که خاصیت آن حساس بودن به این است که یک ویژگی با چه گستردگی و یکنواختی داده‌ها را جدا می‌کند.

۴-۶-۳- مربع کای

مربع کای میزان همبستگی بین دو مقدار را مشخص می‌کند و همانطور که در فصل دوم توضیح داده شد، بررسی می‌کند وقوع یک عبارت خاص و یک کلاس خاص تا چه میزان به یکدیگر وابسته هستند. مقدار مربع کای بالا نشان می‌دهد وقوع عبارت و کلاس مورد نظر، به یکدیگر وابسته هستند.

۴-۷- بررسی سابقه گفتگو

این پایان‌نامه محدودیتی برروی بررسی سابقه مکالمه اعمال نمی‌کند و تمام نوبت‌های قبلی را در پاسخ فعلی تاثیر می‌دهد. بررسی سابقه گفتگو در دسته‌بندی پاسخ به دو شکل وابستگی زمانی و وابستگی ساختاری انجام می‌شود.

۴-۷-۱- وابستگی زمانی

در این روش با استفاده از یک ضریب، عبارت‌های گفته‌های قبلی مکالمه در دسته‌بندی پاسخ، تاثیر داده می‌شوند. ضریب انتخابی مقداری بین صفر و یک خواهد داشت. این ضریب گفته‌هایی که به پاسخ

نزدیک‌ترند را با ارزش بیشتری در دسته‌بندی پاسخ، تاثیر می‌دهند. شکل (۴-۱) مثالی از نحوه استفاده گفته‌های قبلی را در پاسخ فعلی نمایش می‌دهد.

شکل (۴-۱): نحوه تاثیرگذاری نوبت‌های پیشین در پاسخ فعلی

۴-۷-۲- وابستگی ساختاری

منبع [۱۳] تاثیر گفتگوی پیشین را توسط نوع عواطف بکار رفته، بررسی می‌کند و محتویات درون هر نوبت را در دسته‌بندی تاثیر نمی‌دهد. این منبع اندازه حافظه را ثابت در نظر می‌گیرد؛ به عبارت دیگر تنها تعداد نوبت‌های خاصی را در پاسخ تاثیر می‌دهد. در این پایان‌نامه، از این روش برای تاثیر گفتگوی پیشین در دسته‌بندی پاسخ استفاده شده است. با این تفاوت که اندازه حافظه محدود به تعداد نوبت‌ها نیست و کلیه نوبت‌های قبلی در دسته‌بندی تاثیر داده می‌شوند. همچین تاثیرگذاری نوبت به صورت پرسش و نوع پاسخ، خواهد بود. در واقع این روش ارتباط بین کلیه گفته‌ها را تاثیر می‌دهد. سیستم پرسش و پاسخ تعاملی این ارتباط‌ها را به صورت یک گراف در نظر می‌گیرد. به عنوان مثال، اگر پاسخ سوالی خنثی بود، برای ارائه جمله بعدی چندین حالت پیش می‌آید. هنگامی که با ارائه یک پرسش، پاسخ منفی باشد، باید حالت‌های دیگر را در نظر گرفت.

۴-۸- روش‌های ارزیابی

برای مشاهده اثربخشی روش پیشنهاد شده، از روش درصد دقت برای ارزیابی روش‌های مختلف استفاده می‌شود. معیارهای ارزیابی به صورت گسترده در طبقه بندی متن بکار می‌روند عبارتند از

صحت[1]، حساسیت[2]، معیار-F[3] و دقت[4]. نحوه محاسبه این معیارها در رابطه (۴-۳) تا (۴-۴) آورده شده است.

$$percision = \frac{tp}{tp + fp} \quad (۴-۵)$$

$$recall = \frac{tp}{tp + fn} \quad (۴-۶)$$

$$f\ _measure = 2 \times \frac{percision \times recall}{percision + recall} \quad (۴-۷)$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (۴-۸)$$

tp، عناصری از کلاس مورد نظر هستند که در کلاس مورد نظر قرار گرفته‌اند؛

tn، عناصری از کلاس‌های دیگر هستند که در کلاس مورد نظر، قرار نگرفته‌اند؛

fp، عناصری از کلاس‌های دیگر هستند که به عنوان کلاس مورد نظر پیش‌بینی شده‌اند؛

fn، عناصری از کلاس مورد نظر هستند که اشتباه در دسته‌های دیگر پیش‌بینی شده‌اند.

۴-۹- جمع‌بندی

در این فصل به معرفی و تفسیر روش پیشنهادی برای عملیات دسته‌بندی قطبی پاسخ‌ها در یک سیستم پرسش و پاسخ تعاملی پرداخته شد. در ابتدا ویژگی‌هایی که برای هدف انتخاب شده بودند، معرفی شدند. در مرحله بعد علت استفاده از روش‌های دسته‌بندی انتخابی شرح داده شد. برای انتخاب عبارت‌های موثر در دسته‌بندی قطبی از روش‌های استخراج اطلاعات استفاده شد و نحوه عملکرد هر یک از آن‌ها توضیح داده شد. این عملیات پاسخ‌ها را بدون در نظر گرفتن گذشته آن‌ها، دسته‌بندی می‌کنند. برای بررسی تاثیر گفته‌های پیشین دو روش وابستگی زمانی و وابستگی ساختاری معرفی شد.

---

[1] Precision

[2] Recall

[3] F-measure

[4] Accuracy

فصل پنجم:

ارزیابی روش‌های پیشنهادی و بررسی نتایج

۵-۱- مقدمه

در فصل چهارم روش‌هایی برای بهبود دسته‌بندی قطبی پاسخ‌ها در یک سیستم پرسش و پاسخ تعاملی مطرح شد. در این فصل به بررسی و ارزیابی روش‌های معرفی شده، پرداخته می‌شود. در مرحله اول، پایگاه داده مورد استفاده برای دسته‌بندی پاسخ، معرفی می‌شود. سپس روش‌های پیش‌پردازش آزمایش می‌شوند و روش‌های موثر انتخاب می‌شوند. در ادامه با استفاده از روش‌های کاهش ویژگی، عبارت‌های مناسب در دسته‌بندی قطبی پاسخ‌ها، استخراج می‌شوند. این مراحل، عملیات دسته‌بندی قطبی پاسخ‌ها را بدون در نظر گرفتن سابقه گفتگو، تکمیل می‌کنند. برای بررسی تاثیر گفتگوی پیشین در دسته‌بندی پاسخ، دو روش وابستگی زمانی و وابستگی ساختاری آزمایش می‌شوند و با نتایج دسته‌بندی پاسخ‌ها بدون در نظر گرفتن گذشته آنها، مقایسه می‌شود. معیار ارزیابی مورد نظر، درصد دقت می‌باشد.

۵-۲- پایگاه داده

هدف این پژوهش، دسته‌بندی قطبی پاسخ‌های یک سیستم پرسش و پاسخ تعاملی است. بدلیل اینکه هیچ پایگاه داده معتبری در این زمینه وجود ندارد، از مکالمه‌های استاندارد برای این هدف استفاده شده است. تعداد کل مکالمه‌هایی که جمع‌آوری شده‌اند برابر ۲۴۸۵ است. برای هدف دسته‌بندی پاسخ‌ها، ۱۴۴۳ مکالمه انگلیسی انتخاب شدند که در مجموع شامل ۶۹۳۰ پاسخ مثبت، ۲۳۱۹ پاسخ منفی و ۱۱۲۲ پاسخ خنثی هستند. هیچ مکالمه‌ای تنها از یک نوع خاص پاسخ تشکیل نشده است و در هر مکالمه حداقل دو نوع پاسخ وجود دارد. برای توازن در تعداد داده‌های هر دسته، ۱۱۲۲ پاسخ به صورت تصادفی از هر دسته انتخاب شدند. از بین این پاسخ‌ها، برای هردسته ۷۴۸ پاسخ برای آموزش و ۳۷۴ پاسخ برای آزمون درنظر گرفته شده‌اند. این پاسخ‌ها بگونه‌ای هستند که در یک مکالمه هر دو بر چسب آموزش و آزمون وجود نداشته باشد. زیرا در صورت تداخل، هنگامی که تاثیر سابقه مکالمه

بررسی می‌شود، از داده‌هایی استفاده می‌گردد که قبلا آموزش دیده‌اند. درکل ۹۳۷ مکالمه برای آموزش و ۵۰۷ مکالمه برای آزمون استفاده شدند که در مجموعه داده‌های آموزش ۲۲۴۴ پاسخ و در مجموعه داده‌های آزمون، ۱۱۲۲ پاسخ موجود است.

۵-۳- بررسی روش‌های پیش‌پردازش

در این بخش، روش مناسب برای پیش‌پردازش داده‌ها انتخاب می‌شود. برای ارزیابی میزان موثر بودن علائم نگارشی دو روش وجود دارد. در روش اول کلمه‌ها با علائم نگارشی آنها بررسی می‌شوند. در این حالت تعداد ویژگی‌ها در unigram و bigram افزایش می‌یابد. علت این امر، عدم تشخیص ویژگی‌های یکسان است. به عنوان مثال دو ویژگی «know» و«know?» را متفاوت در نظر می‌گیرد. این کار باعث می‌شود دقت کاهش پیدا کند. در روش دوم علائم نگارشی به صورت جداگانه در نظر گرفته شود. در این حالت نیز تعداد ویژگی‌ها افزایش می‌یابد؛ ولی تغییری در دقت ایجاد نمی‌شود. عدم یکسان نمودن شکل‌های مختلف حروف نیز مشکل روش اول را ایجاد می‌کند. به عنوان مثال دو کلمه «YES» و «Yes» را یکسان در نظر می‌گیرد. این کار علاوه بر افزایش تعداد ویژگی‌ها، دقت روش‌های دسته-بندی را نیز کاهش می‌دهد.در صورتی که حروف تکراری حذف نشوند، هر کلمه به صورت یک ویژگی در نظر گرفته می‌شود. به عنوان مثال «noooooooooo» ممکن است در کل متن‌های آموزشی تنها یکبار تکرار شود. زیرا تعداد حروف تکراری در یک کلمه متفاوت است. عدم حذف حروف تکراری نسبت به دو روش دیگر بیشتر دقت را کاهش می‌دهد. بنابراین در پیش‌پردازش اولیه، در ابتدا داده‌ها با حذف علائم نگارشی، یکسان نمودن شکل‌های مختلف حروف و حذف حالت‌های نامتعادل، پیش-پردازش می‌شوند. در مراحل بعدی روش‌های ریشه‌یابی و حذف کلمات اضافی نیز به پیش‌پردازش اولیه اضافه می‌شوند و نتایج بررسی خواهند شد.

۵-۳-۱- پیش‌پردازش اولیه

همانطور که گفته شد، پیش‌پردازش اولیه شامل حذف علامت‌های نگارشی، یکسان نمودن شکل‌های مختلف حروف و حذف حالت‌های نامتعادل می‌شود. جدول (۵-۱) نتایج دسته‌بندی‌های مختلف را بر

روی unigram نشان می‌دهد. داده‌ها در این جدول توسط پیش‌پردازش اولیه اصلاح شدند. برای راحتی توضیحات، از حالت اختصار اطلاعات متقابل نقطه به نقطه یا به بیانی دیگر PMI استفاده می-شود.

جدول (۵-۱): پیش‌پردازش اولیه و unigram

	فرکانس	حضور	TF-IDF	ویژگی احتمالی	PMI
ماشین بردار پشتیبان (چند جمله‌ای)	۷۵/۹۳	۷۵/۲۲	۷۵/۲۲	۷۶/۰۲	۷۶/۰۲
ماشین بردار پشتیبان (خطی)	۷۵/۹۳	۷۵/۲۲	۶۷/۰۲	۷۷/۵۴	۷۷/۵۵
ماشین بردار پشتیبان (RBF)	۴۵/۸۱	۶۰/۶۹	۴۸/۲۱	۵۷/۸۴	۶۱/۴۹
بیزین ساده	۶۴/۹۷	۶۶/۱۳	۶۶/۱۳	۶۴/۹۷	۶۴/۹۷
درخت تصمیم	۷۰/۶۷	۷۰/۴۹	۷۰/۴۹	۷۰/۶۷	۷۰/۶۷
Bagging (درخت تصمیم)	۷۴/۵۰	۷۴/۴۲	۷۴/۴۲	۷۴/۵۰	۷۴/۵۰

با در نظر گرفتن unigram، ۲۲۰۴ کلمه از مجموعه داده‌های آموزشی استخراج شدند. با توجه به جدول (۵-۱)، ماشین بردار پشتیبان(خطی) و PMI بهترین درصد دقت را با مقدار ۷۷/۵۴ بدست آورد. در میان کلیه روش‌های دسته‌بندی، ماشین بردار پشتیبان (چندجمله‌ای و خطی) درصد دقت بالاتری دارد. در ماشین بردار پشتیبان (چندجمله‌ای) نتایج ویژگی‌ها بسیار نزدیک بهم هستند. با این حال توابع ویژگی احتمالی و PMI با مقدار ٪۷۶/۰۶، بالاترین درصد دقت را در این روش بدست آوردند. برخلاف ماشین بردار پشتیبان (چندجمله‌ای)، درصد دقت ویژگی‌های مختلف در ماشین بردار پشتیبان (خطی)، با یکدیگر متفاوت هستند. ویژگی TF-IDF در این روش، مناسب نیست. زیرا نسبت به سایر ویژگی‌ها، درصد دقت داده‌های آزمون را به شدت کاهش می‌دهد. ویژگی احتمالی و PMI در ماشین بردار پشتیبان (خطی) با دقت ۷۷/۵۴ و ۷۷/۵۵ بیشترین درصد دقت را دربین ویژگی‌ها و روش‌های دسته‌بندی بدست آوردند. بدترین دقت در تمام ویژگی‌ها را ماشین بردار پشتیبان با هسته تابع پایه شعاعی بدست آورد. بیزین ساده با تمام ویژگی‌ها عملکرد ضعیفی دارد. در این روش دسته-بندی، حضور عبارت و TF-IDF کلاسیک با مقدار ۶۶/۱۳ بیشترین درصد دقت را در بین ویژگی‌ها، بدست آوردند. در درخت تصمیم، نتایج درصد دقت ویژگی‌های مختلف به یکدیگر نزدیک هستند. فرکانس عبارت، ویژگی احتمالی و PMI با مقدار ۷۰/۶۷ بیشترین درصد دقت را در درخت تصمیم بدست آوردند. الگوریتم Bagging (درخت تصمیم) مشابه درخت تصمیم، در فرکانس عبارت، ویژگی

احتمالی و PMI با مقدار ۷۴/۵۰٪ بیشترین درصد دقت را بدست آوردند. درصد دقت الگوریتم Bagging (درخت تصمیم)، نسبت به استفاده از درخت تصمیم بهتر بوده است. همانطور که در فصل چهارم توضیح داده شد، الگوریتم Bagging داده‌های آموزش را با جایگزینی، به زیر مجموعه‌هایی تبدیل می‌کند. سپس بر روی هر زیر مجموعه یک مدل دسته‌بندی را اعمال می‌کند و نتایج را با یکدیگر ترکیب می‌کند. در این آزمایش از درخت تصمیم برای آموزش زیرمجموعه‌ها استفاده شده است. همانطور که مشاهده می‌شود، استفاده از چند درخت تصمیم و ترکیب نتایج آنها، درصد دقت بالاتری را نسبت به استفاده از درخت تصمیم برروی کلیه داده‌های آموزش، بدست می‌آورد.

همانطور که در فصل چهارم توضیح داده شد، فرضیه این پژوهش بر این مبنا است که استفاده از bigram دقت بالاتری را نسبت به unigram بدست می‌آورد. جدول(۵-۲) دقت دسته‌بندی‌های مختلف را با bigram نشان می‌دهد که در آن داده‌ها توسط پیش‌پردازش اولیه اصلاح شدند.عبارت‌های استخراج شده توسط bigram برابر ۹۵۸۲ عبارت می‌باشند. با توجه به تعداد زیاد ویژگی‌ها و تعداد داده‌های آموزشی محدود، این احتمال وجود دارد که پارامترهای روش‌های دسته‌بندی، به خوبی تنظیم نشوند و مشکل کم برازش[۱] بوجود آید.

جدول (۵-۲): پیش‌پردازش اولیه و bigram

	فرکانس	حضور	TF-IDF	ویژگی احتمالی	PMI
ماشین بردار پشتیبان (چند جمله‌ای)	۷۲/۳۷	۷۲/۳۷	۷۲/۳۷	۷۲/۳۷	۷۲/۳۷
ماشین بردار پشتیبان (خطی)	۷۲/۴۵	۷۲/۳۷	۶۵/۴۱	۷۲/۹۰	۷۲/۷۲
ماشین بردار پشتیبان (RBF)	۵۴/۸۱	۵۴/۴۵	۵۹/۵۳	۴۴/۴۷	۵۰/۶۲
بیزین ساده	۶۳/۰۱	۶۳/۰۱	۶۳/۰۱	۶۳/۰۱	۶۳/۰۱
درخت تصمیم	۵۶/۶۸	۵۶/۶۸	۵۶/۶۸	۵۶/۶۸	۵۶/۶۸
Bagging (درخت تصمیم)	۵۷/۴۸	۵۷/۴۸	۵۷/۴۸	۵۷/۴۸	۵۷/۴۸

جدول (۵-۲) نتایج درصد دقت دسته‌بندی‌های مختلف را برروی داده‌های آزمون نمایش‌می‌دهد که توسط پیش‌پردازش اولیه اصلاح و از bigram استفاده شده است. به طور کلی نتایج نسبت به unigram، کاهش یافته است. علت این مشکل می‌تواند نویزی بودن بعضی از عبارت‌ها و یا تعداد زیاد وعبارت‌ها نسبت به داده‌های آموزش، در نظر گرفته شود. مشابه آزمایش‌های قبل، ماشین بردار

[۱] Underfitting

پشتیبان(خطی و چندجمله‌ای) نسبت به سایر روش‌های دسته‌بندی، درصد دقت بالاتری دارد. بهترین درصد دقت توسط ویژگی احتمالی و ماشین بردار پشتیبان(خطی) با مقدار ۷۲/۹۰٪ بدست آمد. همانطور که مشاهده می‌شود، دقت ویژگی‌های مختلف در روش‌های دسته‌بندی ماشین بردار پشتیبان(چند جمله‌ای)، درخت تصمیم و الگوریتم Bagging (درخت تصمیم)، مشابه یکدیگر هستند. دقت استفاده از ویژگی‌های مختلف در ماشین بردار پشتیبان(خطی)، با یکدیگر متفاوت است. ویژگی TF-IDF کلاسیک با مقدار۶۵/۴۱٪ کمترین دقت را در ماشین بردار پشتیبان(خطی) را داشت. در مقایسه با جدول (۵-۱)، درخت تصمیم، کمترین دقت دسته‌بندی را بدست آورد. همچنین الگوریتم Bagging (درخت تصمیم)، دقت بالاتری را نسبت به استفاده از درخت تصمیم بدست آورد. بدترین دقت در تمام ویژگی‌ها را ماشین بردار پشتیبان با هسته تابع پایه شعاعی بدست آورد.

۵-۳-۲- ترکیب پیش‌پردازش اولیه با ریشه‌یابی

در این آزمایش علاوه بر پیش‌پردازش اولیه، کلمه‌های پاسخ‌ها توسط ریشه‌یاب وکا، به حالت پایه خود تبدیل می‌شوند. جدول‌های (۵-۳) و (۵-۴) نتایج ارزیابی‌های مختلف را بر روی داده‌های آزمون نمایش می‌دهند. در این بخش جدول (۵-۳) نتایج را با unigram و جدول (۵-۴) نتایج را با bigram نشان نشان می‌دهد.

جدول (۵-۳): ترکیب پیش‌پردازش اولیه با ریشه‌یابی و unigram

	فرکانس	حضور	TF-IDF	ویژگی احتمالی	PMI
ماشین بردار پشتیبان (چند جمله‌ای)	۷۵/۲۲	۷۵/۱۳	۷۵/۱۳	۷۵/۲۲	۷۵/۲۲
ماشین بردار پشتیبان (خطی)	۷۵/۰۴	۷۵/۱۳	۶۷/۶۴	۷۷/۴۵	۷۵/۷۵
ماشین بردار پشتیبان (RBF)	۴۸/۱۲	۶۲/۲۱	۴۶/۵۲	۵۳/۲۹	۶۱/۶۷
بیزین ساده	۶۷/۲۰	۶۵/۳۲	۶۵/۳۲	۶۷/۲۰	۶۷/۲۰
درخت تصمیم	۷۲/۹۹	۷۲/۷۲	۷۲/۷۲	۷۲/۹۹	۷۲/۹۹
Bagging (درخت تصمیم)	۷۴/۰۶	۷۴/۷۷	۷۴/۷۷	۷۴/۰۶	۷۴/۰۶

با تبدیل کلمه‌ها به حالت پایه با استفاده از ریشه‌یاب وکا، اندازه مجموعه ویژگی از ۲۲۰۴ کلمه به ۱۶۸۷ کاهش می‌یابد. جدول(۵-۳) دقت را بر روی داده‌های آزمون با استفاده از پیش‌پردازش اولیه و ریشه‌یابی کلمه‌ها نشان می‌دهد. بهترین دقت توسط ماشین بردار پشتیبان (خطی) و ویژگی احتمالی با مقدار ۷۷٪/۴۵ بدست آمد. به طور کلی ماشین بردار پشتیبان(خطی و چندجمله‌ای) در اکثر

ویژگی‌ها دقت بالاتری بدست آورد و ضعیف‌ترین دقت را بیزین ساده داشت. ویژگی احتمالی و PMI در اکثر روش‌های دسته‌بندی دقت بالاتری داشتند. دقت ماشین بردار پشتیبان(خطی و چندجمله‌ای) و الگوریتم Bagging در اکثر ویژگی‌ها نسبت به جدول (۵-۱) بصورت جزئی کاهش یافت. در مقابل عملکرد بیزین ساده و درخت تصمیم با ترکیب پیش‌پردازش و ریشه‌یابی بهبود یافت. بدترین دقت در تمام ویژگی¬ها را ماشین بردار پشتیبان با هسته تابع پایه شعاعی بدست آورد. بدلیل اینکه اندازه ویژگی‌ها به طور چشمگیری در این روش پیش‌پردازش، کاهش یافت، می‌توان از این روش به عنوان روش پیش‌پردازش برای باقی آزمایش‌ها، استفاده کرد. اما این روش به صورت جزئی دقت بعضی از روش‌های دسته‌بندی را کاهش می‌دهد که می‌توان این کاهش جزئی را نادیده گرفت.

با ترکیب ریشه‌یابی و پیش‌پردازش اولیه مشاهده می‌شود تعداد عبارت‌های bigram از ۹۵۸۲ به ۹۱۶۴ ویژگی کاهش می‌یابد. دقت دسته‌بندی‌های مختلف بر روی bigram توسط این پیش‌پردازش در جدول (۵-۴) نمایش داده می‌شود.

جدول (۵-۴): ترکیب پیش‌پردازش اولیه با ریشه‌یابی و bigram

	فرکانس	حضور	TF-IDF	ویژگی احتمالی	PMI
ماشین بردار پشتیبان (چند جمله‌ای)	۷۲/۱۰	۷۲/۳۷	۷۲/۳۷	۷۲/۱۰	۷۲/۱۰
ماشین بردار پشتیبان (خطی)	۷۲/۹۹	۷۲/۳۷	۷۲/۳۷	۷۵/۵۴	۷۵/۲۸
ماشین بردار پشتیبان (RBF)	۵۳/۸۳	۵۳/۹۲	۵۹/۴۴	۴۳/۷۶	۴۹/۰۱
بیزین ساده	۶۳/۳۶	۶۲/۲۱	۶۲/۲۱	۶۳/۳۶	۶۳/۳۶
درخت تصمیم	۵۶/۵۹	۵۶/۵۹	۵۶/۵۹	۵۶/۵۹	۵۶/۵۹
Bagging (درخت تصمیم)	۵۷/۳۹	۵۷/۳۹	۵۷/۳۹	۵۷/۳۹	۵۷/۳۹

در همه جداول ماشین بردار پشتیبان با هسته تابع شعاعی عملکرد بسیار پایینی دارد. با مقایسه جدول (۵-۴) با جدول (۵-۲) مشاهده می‌شود درصد دقت ماشین بردار پشتیبان (چندجمله‌ای) در ویژگی‌های فرکانس عبارت، ویژگی احتمالی و PMI به صورت بسیار جزئی کاهش یافته است. در مقابل درصد دقت ماشین بردار پشتیبان (خطی) و بیزین ساده بهبود یافته است. درصد دقت درخت تصمیم و الگوریتم Bagging (درخت تصمیم) نیز کاهش یافته است و در تمامی ویژگی‌ها ثابت هستند. ویژگی احتمال و PMI نسبت به سایر ویژگی در اکثر روش‌های دسته‌بندی، دقت بالاتری

دارند. بهترین دقت توسط ماشین بردار پشتیبان(خطی) و ویژگی احتمالی بدست آمد و این دقت برابر ۷۵/۵۴٪ می‌باشد.

با بررسی کلی جدول‌های (۵-۳) تا (۵-۴) این نتیجه حاصل می‌شود که استفاده از ریشه‌یابی برای پیش‌پردازش، هرچند موجب کاهش اندازه مجموعه ویژگی می‌شود. اما به صورت جزئی دقت الگوریتم‌های دسته‌بندی را نیز کاهش می‌دهد. ولی با توجه به اینکه این کاهش دقت جزئی است و این تعداد ویژگی می‌تواند سرعت اجرای الگوریتم‌های زمانبر مانند الگوریتم Bagging (درخت تصمیم) را کاهش دهد، بنابراین بهتر است در آزمایش‌های بعدی از این روش به عنوان پیش‌پردازش داده‌ها، استفاده شود.

۵-۳-۳-بررسی پیش‌پردازش با حذف کلمات اضافی

در این روش، داده‌ها علاوه بر اینکه توسط پیش‌پردازش اولیه اصلاح می‌شوند؛ کلمه‌های اضافی موجود در آن‌ها نیز، حذف می‌شود. تاثیر این روش پیش‌پردازش در جدول‌های (۵-۵) و(۵-۶) نمایش داده می‌شود.

با حذف کلمات اضافی اندازه unigram به ۲۰۶۰ کلمه کاهش می‌دهد. همانطور که در فصل چهارم توضیح داده شد، از کتابخانه استندفورد برای شناسایی این نوع کلمه‌ها استفاده شده است.

جدول (۵-۵): ترکیب پیش‌پردازش اولیه با حذف کلمات اضافی و unigram

	فرکانس	حضور	TF-IDF	ویژگی احتمالی	PMI
ماشین بردار پشتیبان (چند جمله‌ای)	۶۹/۷۹	۶۹/۰۷	۶۹/۰۷	۶۹/۷۹	۶۹/۷۹
ماشین بردار پشتیبان (خطی)	۶۹/۵۱	۶۹/۰۷	۶۳/۱۱	۶۹/۵۲	۶۹/۵۲
ماشین بردار پشتیبان (RBF)	۳۳/۷۲	۳۳/۷۲	۵۴/۱۹	۳۳/۷۲	۴۸/۸۲
بیزین ساده	۵۹/۷۹	۵۹/۶۹	۵۹/۶۹	۵۹/۷۹	۵۹/۷۹
درخت تصمیم	۵۶/۶۲	۵۶/۶۲	۵۹/۶۲	۵۹/۶۲	۵۹/۶۲
Bagging (درخت تصمیم)	۵۷/۶۱	۵۷/۷۰	۵۷/۷۰	۵۷/۶۱	۵۷/۶۱

جدول (۵-۵) درصد دقت داده‌های آزمون را با استفاده از بر روی unigram با پیش‌پردازش اولیه و حذف کلمه‌های اضافی نشان می‌دهد. مشابه تمام آزمایش‌های قبلی ماشین بردار پشتیبان (چندجمله‌ای و خطی) همچنان نسبت به سایر روش‌ها، دقت بالاتری دارد. پایین‌ترین درصد دقت را درخت تصمیم بدست آورد. ویژگی احتمالی و PMI نیز همانند آزمایش‌های قبلی، نسبت به سایر

ویژگی‌ها درصد دقت بالاتری را در اکثر روش‌های دسته‌بندی کسب کردند. درخت تصمیم با استفاده از ویژگی‌هایی مانند TF-IDF کلاسیک، ویژگی احتمالی و PMI، دقت بالاتری نسبت به الگوریتم Bagging بدست آورد. حضور عبارت و TF-IDF در الگوریتم Bagging (درخت تصمیم)، به صورت جزئی دقت بالاتری را نسبت به ویژگی احتمالی و PMI بدست آوردند. در مقایسه با جدول (۵-۱)، مشاهده می‌شود دقت روش‌های مختلف به طرز آشکاری کاهش پیدا کرد. همچنین با انتخاب تابع پایه شعاعی به عنوان هسته ماشین بردار پشتیبان، دقت بششدت کاهش می‌یابد.

در صورتی که bigram به عنوان انتخاب شوند، با حذف کلمات اضافی، اندازه bigram از ۹۵۸۲ عبارت به ۹۴۰۵ عبارت کاهش یافت. جدول (۵-۶) نتایج دسته‌بندی‌های مختلف را توسط این روش پیش-پردازش نشان می‌دهد.

جدول (۵-۶): ترکیب پیش‌پردازش اولیه با حذف کلمات اضافی و bigram

	فرکانس	حضور	TF-IDF	ویژگی احتمالی	PMI
ماشین بردار پشتیبان (چند جمله‌ای)	۵۵	۵۴/۷۳	۵۴/۷۳	۵۵	۵۵
ماشین بردار پشتیبان (خطی)	۵۵/۰۹	۵۴/۸۲	۵۴/۲۲	۵۵/۳۶	۵۵/۱۸
	۳۳/۷۲	۳۳/۷۲	۳۷/۸۷	۷۲/۳۳	۳۳/۷۲
بیزین ساده	۵۱/۴۸	۵۱/۳۹	۵۱/۳۹	۵۱/۴۸	۵۱/۴۸
درخت تصمیم	۳۳/۷۲	۳۳/۷۲	۳۳/۷۲	۳۳/۷۲	۳۳/۷۲
Bagging (درخت تصمیم)	۳۴/۰۸	۳۴/۰۸	۳۴/۰۸	۳۴/۰۸	۳۴/۰۸

جدول (۵-۶) درصد دقت را بر روی مجموعه داده‌های آزمون نشان می‌دهند که توسط پردازش اولیه و حذف کلمات اضافی اصلاح شدند. بدلیل اینکه سه روش ماشین بردار پشتیبان (RBF) درخت تصمیم و الگوریتم Bagging (درخت تصمیم) اکثر داده‌ها را جزء کلاس مثبت در نظر گرفتند، دقت این دو الگوریتم برای داده‌های آزمون بسیار پایین است. ماشین بردار پشتیبان (خطی و چندجمله‌ای) نسبت به سایر روش‌ها درصد دقت بالاتری دارد. بدترین دقت را درخت تصمیم داشت زیرا این روش، کلیه داده‌های آزمون را در دسته داده‌های مثبت قرار داد. در مقایسه با استفاده از پیش پردازش اولیه می-توان به این نتیجه رسید که استفاده از حذف کلمه‌های اضافی روش نامناسب برای دسته‌بندی پاسخ‌ها است.

در بخش (۵-۳) به بررسی روش‌های متداول پیش‌پردازش در پردازش متن، پرداخته شد. ارزیابی‌ها بر روی unigram و bigram انجام شد آزمایش شدند. ویژگی احتمالی و اطلاعات متقابل نقطه به نقطه، در بین سایر ویژگی‌ها، دقت بهتری داشتند. در میان روش‌های دسته‌بندی، ماشین بردار پشتیبان (چندجمله‌ای و خطی)، عملکرد بهتری داشت. در مقابل ماشین بردار پشتیبان (RBF) بسیار ضعیف عمل می‌کرد. به دلیل تفاوت دقت زیاد آن با سایر روش‌های دسته‌بندی، در آزمایش‌های بعدی این روش حذف می‌شود. در میان تمامی روش‌های پیش‌پردازش، پیش‌پردازش اولیه تقریبا نسبت به سایر روش‌ها، درصد دقت بالاتری داشت. همچنین ترکیب پیش‌پردازش اولیه و ریشه‌یابی کمترین تعداد عبارت را داشت؛ اما درصد دقت، در بعضی از روش‌های دسته‌بندی نسبت به پیش‌پردازش اولیه، کمتر بود. ترکیب پیش‌پردازش اولیه و حذف کلمه‌های اضافی، بدترین درصد دقت را در تمامی روش‌های دسته‌بندی بدست آورد. علت این پدیده، وجود کلمات نفی در مجموعه کلمات اضافی بود. در ادامه آزمایش‌ها، نمونه‌ها توسط پیش‌پردازش اولیه و ریشه‌یابی اصلاح می‌شوند.

۵-۴- بررسی روش‌های کاهش ویژگی

یکی از مشکلاتی که در آزمایش‌های unigram و bigram وجود دارد، اندازه آن‌ها است. ابعاد بزرگ ویژگی‌ها باعث می‌شود روش‌های دسته‌بندی نتوانند به خوبی آموزش ببینند. همچنین پارامترهای مربوط به آن‌ها به خوبی تنظیم نشود. همانطور که در فصل چهارم توضیح داده شد، برای کاهش ویژگی از روش‌های بهره اطلاعات، نرخ بهره و مربع کای استفاده می‌شود. داده‌ها توسط پیش‌پردازش اولیه و ریشه‌یابی اصلاح شدند. زیرا این روش پیش‌پردازش توانست کمترین تعداد ویژگی را تولید کند. با توجه به اینکه ویژگی احتمالی و اطلاعات متقابل نقطه به نقطه نسبت به سایر ویژگی‌ها، عملکرد بهتری داشتند، روش‌های کاهش ویژگی بر روی آنها آزمایش می‌شوند. آزمایش‌های مربوط به این قسمت در سه مرحله ارزیابی می‌شوند. در مرحله اول روش‌های کاهش ویژگی بر روی unigram اعمال می‌شود؛ در مرحله دوم این روش‌ها بر روی bigram آزمایش می‌شوند و در مرحله سوم، از ترکیب unigram و bigram استفاده می‌شود.

۵-۴-۱-بررسی روش‌های کاهش ویژگی بر روی unigram

بهره اطلاعات، نرخ بهره و مربع کای به طور کلی ۱۴۳ کلمه از بین ۱۶۸۷ کلمه در هر دو تابع وزن-دهی اطلاعات متقابل نقطه به نقطه و ویژگی احتمالی استخراج کردند.

جدول (۵-۷): تأثیر روش‌های کاهش ویژگی بر روی unigram و ویژگی احتمالی

	بهره اطلاعات	نرخ بهره	مربع کای
ماشین بردار پشتیبان (چند جمله‌ای)	۷۵/۳۱	۷۵/۳۱	۷۵/۳۱
ماشین بردار پشتیبان (خطی)	۷۵/۶۶	۷۵/۶۶	۷۵/۶۶
بیزین ساده	۶۸/۸۰	۶۸/۸۰	۶۸/۸۰
درخت تصمیم	۷۲/۹۰	۷۲/۹۰	۷۲/۹۰
Bagging (درخت تصمیم)	۷۳/۱۷	۷۳/۱۷	۷۳/۱۷

همانطور که در جدول (۵-۷) مشاهده می‌شود، درصد دقت هر سه روش کاهش ویژگی مشابه است. علت این امر این است که هر سه روش توانستند کلمات یکسانی را استخراج کنند. بهترین دقت را ماشین بردار پشتیبان (خطی) و بدترین را بیزین ساده بدست آورد. در مقایسه با جدول (۵-۳)، مشاهده می‌شود که درصد دقت روش دسته‌بندی ماشین بردار پشتیبان(چندجمله‌ای) از ۷۵/۱۳ به ۷۵٪/۳۱ و بیزین ساده از ۶۵/۳۲٪ به ۶۸/۸۰٪ بهبود یافته است. در مقابل درصد دقت ماشین بردار پشتیبان (خطی) از ۷۷/۴۵٪ به ۷۵/۶۶٪، درخت تصمیم از ۷۲/۹۹٪ به ۷۲/۹۰٪ و الگوریتم Bagging (درخت تصمیم) از ۷۴/۰۶٪ به ۷۳/۱۷٪ کاهش یافته است. همانطور که مشاهده می‌شود، روش‌های کاهش ویژگی دقت اکثر روش‌های دسته‌بندی را به صورت جزئی کاهش می‌دهد؛ ولی می‌توانند عبارت‌های ارزشمند در دسته‌بندی را استخراج کنند که این عمل موجب می‌شود اندازه مجموعه ویژگی به شدت کاهش یابد.

جدول (۵-۸): تأثیر روش‌های کاهش ویژگی بر روی unigram و PMI

	بهره اطلاعات	نرخ بهره	مربع کای
ماشین بردار پشتیبان (چند جمله‌ای)	۷۵/۳۱	۷۵/۳۱	۷۵/۳۱
ماشین بردار پشتیبان (خطی)	۷۵/۴۹	۷۵/۴۹	۷۵/۴۹
بیزین ساده	۶۸/۸۰	۶۸/۸۰	۶۸/۸۰
درخت تصمیم	۷۲/۹۰	۷۲/۹۰	۷۲/۹۰
Bagging (درخت تصمیم)	۷۳/۱۷	۷۳/۱۷	۷۳/۱۷

جدول (۵-۸) درصد دقت ویژگی PMI را توسط روش‌های کاهش ویژگی، نشان می‌دهد. همانطور که مشاهده می‌شود، ماشین‌بردار پشتیبان (خطی و چندجمله‌ای) نسبت به سایر روش‌های دسته‌بندی دقت بالاتری دارد. با مقایسه دو جدول (۵-۷) و (۵-۸) این نتیجه حاصل می‌شود که دقت ماشین بردار پشتیبان (خطی) با استفاده از PMI کاهش می‌یابد و دقت ماشین بردار پشتیبان(چند جمله‌ای)، بیزین ساده، درخت تصمیم و الگوریتم Bagging بدون تغییر باقی می‌مانند. در مقایسه با جدول (۵-۳)، دقت ماشین بردار پشتیبان(خطی) از ۷۵/۷۵٪ به ۷۵/۴۹٪، درخت تصمیم از ۷۲/۹۹٪ به ۷۲/۹۰٪ و الگوریتم Bagging (درخت تصمیم) از ۷۴/۰۶٪ به ۷۳/۱۳٪ کاهش می‌یابد. در مقابل ماشین بردار پشتیبان (چند جمله‌ای) از ۷۵/۲۲٪ به ۷۵/۳۱٪ و بیزین ساده از ۶۷/۲۰٪ به ۶۸/۸۰٪ بهبود یافته است. مشابه جدول (۵-۷) استفاده از PMI در unigram موجب می‌شود روش‌های استخراج ویژگی بهره اطلاعات، نرخ بهره و مربع کای کلمه‌های یکسانی را استخراج کنند. بنابراین درصد دقت روش-های دسته‌بندی، در هر سه روش استخرج ویژگی یکسان است.

با بررسی کلی آزمایش‌ها این نتیجه حاصل می‌شود که بهترین دقت در استفاده از مجموعه ویژگی اصلی (بدون کاهش ویژگی)، توسط ماشین بردار پشتیبان (خطی) و ویژگی احتمالی بدست آمد. این مقدار برابر ۷۷/۴۵٪ است که با اعمال روش‌های استخراج ویژگی ۲٪ کاهش یافت و مقدار آن برابر ۷۵/۶۶ رسید. در مقابل بهترین درصد دقت ویژگی PMI، توسط ماشین بردار پشتیبان(خطی) با درصد ۷۵/۷۵٪ بدست آمد که با اعمال روش‌های کاهش ویژگی به صورت خیلی جزئی کاهش پیدا کرد و مقدار آن به ۷۵/۴۹٪ رسید.

۵-۴-۲-بررسی روش‌های کاهش ویژگی بر روی bigram

با اعمال روش‌های کاهش ویژگی اندازه مجموعه ویژگی از ۹۱۶۴ عبارت به ۴۲۸ عبارت کاهش یافت. این اندازه ویژگی موجب می‌شود، سرعت اگوریتم‌های دسته‌بندی افزایش پیدا کند.

جدول (۵-۹): تأثیر روش‌های کاهش ویژگی بر روی bigram و ویژگی احتمالی

	بهره اطلاعات	نرخ بهره	مربع کای
ماشین بردار پشتیبان (چند جمله‌ای)	۶۸/۲۷	۶۸/۲۷	۶۸/۱۸
ماشین بردار پشتیبان (خطی)	۶۸/۴۴	۶۸/۴۴	۶۸/۳۶
بیزین ساده	۶۱/۴۹	۶۱/۴۹	۶۱/۴۹
درخت تصمیم	۵۵/۲۵	۵۵/۲۵	۵۵/۲۵
Bagging (درخت تصمیم)	۵۶/۲۳	۵۶/۲۳	۵۶/۲۳

همانطور که در جدول (۵-۹) مشاهده می‌شود، دقت دسته‌بندی‌های مختلف با استفاده از بهره اطلاعات و نرخ بهره، یکسان است. زیرا دو روش بهره اطلاعات و نرخ بهره، کلمه‌های یکسانی را استخراج کردند. دقت مربع کای در ماشین بردار پشتیبان (خطی و چند جمله‌ای)، نسبت به دو روش دیگر کمتر است. علت این امر این است که مربع کای یک کلمه متفاوت استخراج کرده است. با مقایسه جدول (۵-۹) و (۵-۴) این نتیجه حاصل می‌شود که استفاده از روش‌های کاهش ویژگی، به شدت دقت دسته‌بندی‌های مختلف را بر روی bigram و ویژگی احتمالی کاهش می‌دهد. بیشترین کاهش دقت در هر سه روش کاهش ویژگی، بر روی ماشین بردار پشتیبان (خطی و چندجمله‌ای) رخ داده که مقدار آن را حدود ۴٪ کاهش داده است.

جدول (۵-۱۰): تاثیر روش‌های کاهش ویژگی بر روی bigram و PMI

	بهره اطلاعات	نرخ بهره	مربع کای
ماشین بردار پشتیبان (چند جمله‌ای)	۶۸/۲۷	۶۸/۲۷	۶۸/۱۸
ماشین بردار پشتیبان (خطی)	۶۶/۹۳	۶۶/۹۳	۶۶/۹۳
بیزین ساده	۶۱/۴۹	۶۱/۴۹	۶۱/۴۹
درخت تصمیم	۵۵/۲۵	۵۵/۲۵	۵۵/۲۵
Bagging (درخت تصمیم)	۵۶/۲۳	۵۶/۲۳	۵۶/۲۳

جدول (۵-۱۰) نتایج مربوط به اعمال روش‌های کاهش ویژگی بر روی bigram و PMI را نشان می-دهد. بهترین دقت توسط ماشین بردار پشتیبان (چندجمله‌ای) بدست آمد. مشابه جدول (۵-۹)، با مقایسه دو جدول (۵-۱۰) و (۵-۴) این نتیجه حاصل می‌شود که استفاده از روش‌های کاهش ویژگی بر روی bigram و PMI دقت روش‌های دسته‌بندی مختلف را کاهش می‌دهد. درصد دقت روش‌های دسته‌بندی دو جدول (۵-۹) و (۵-۱۰) مشابه یکدیگر هستند؛ ولی در ماشین بردار پشتیبان(خطی)

در جدول (۵-۹)، دقت بیشتری نسبت به جدول (۵-۱۰) دارد و درصد دقت مربع کای نسبت به بهره اطلاعات و نرخ بهره متفاوت است.

مجموعه ویژگی bigram برای هدف دسته‌بندی پاسخ‌ها، ابعاد بزرگی دارد و این موجب می‌شود، دقت روش‌های دسته‌بندی نسبت به استفاده از unigram، کمتر باشد. طبق فرضیه این پژوهش، استفاده از bigram بایستی دقت بیشتری نسبت به unigram بدست آورد. برای حل مشکل ابعاد، بر روی unigram و bigram روش‌های کاهش ویژگی اعمال شد و نتایج در جدول‌های (۵-۷) تا (۵-۱۰) بررسی شدند. با تحلیل جدول‌های (۵-۷) تا (۵-۱۰) این نتیجه حاصل می‌شود که استفاده از روش-های کاهش ویژگی بر روی bigram، دقت روش‌های مختلف دسته‌بندی را بشدت کاهش می‌دهد. درحالی که اعمال روش‌های کاهش ویژگی بر روی unigram باعث ایجاد تغییرات جزئی می‌شود.

با اعمال روش‌های کاهش ویژگی، دقت بعضی از روش‌های دسته‌بندی به صورت جزئی کاهش می‌یابد. یکی از موارد مهم در یک سیستم پرسش و پاسخ، پردازش سریع است. کاربران سیستم‌هایی را که در زمان کمتری بتوانند به آنها پاسخ دهند را ترجیح می‌دهند. به عنوان مثال بدون کاهش ویژگی، زمان پردازش الگوریتم bagging بر روی ویژگی احتمالی bigramها برابر ۶ دقیقه و ۱۳ ثانیه است. در حالی که با اعمال کاهش ویژگی این زمان به ۸ ثانیه کاهش می‌یابد.

۵-۴-۳-بررسی روش‌های کاهش ویژگی بر روی ترکیب unigram و bigram

با بررسی روش‌های کاهش ویژگی بر روی دو مجموعه ویژگی‌های unigram و bigram بصورت جداگانه، این نتیجه حاصل شد که اعمال روش‌های کاهش ویژگی بر روی bigram موجب کاهش دقت روش‌های دسته‌بندی مختلف می‌شود. در این بخش هر دو مجموعه کاهش یافته استخراج شده در هر سه روش کاهش ویژگی با هم ترکیب می‌شوند و نتایج بررسی خواهد شد.

با ترکیب دو مجموعه ویژگی unigram و bigram کاهش یافته، بهره اطلاعات و نرخ بهره ۵۷۰ عبارت استخراج کردند و مربع کای ۵۶۸ عبارت بدست آورد. ویژگی مورد آزمایش در این قسمت ویژگی احتمالی است.

جدول (۵-۱۱): ترکیب unigram و bigram کاهش یافته و ویژگی احتمالی

	بهره اطلاعات	نرخ بهره	مربع کای
ماشین بردار پشتیبان (چند جمله‌ای)	۷۵/۵۷	۷۵/۵۷	۷۵/۶۶
ماشین بردار پشتیبان (خطی)	۷۶/۱۱	۷۶/۱۱	۷۶/۱۱
بیزین ساده	۶۸/۲۷	۶۸/۲۷	۶۸/۲۷
درخت تصمیم	۷۳/۴۴	۷۳/۴۴	۷۳/۴۴
Bagging (درخت تصمیم)	۷۳/۷۰	۷۳/۷۰	۷۳/۷۰

جدول (۵-۱۱) تاثیر ترکیب n-gram های کاهش یافته را بر روی دقت دسته‌بندی‌های مختلف، نشان می‌دهد. بهترین دقت توسط ماشین بردار پشتیبان (خطی) با مقدار ۷۶/۱۱٪ بدست آمد. مشابه آزمایش‌های قبلی، دو روش بهره اطلاعات و نرخ بهره ویژگی‌های یکسانی دارند و دقت روش‌های دسته‌بندی توسط این دو روش، مشابه است. مربع کای در ماشین بردار پشتیبان(چند جمله‌ای) دقت بالاتری نسبت به دو روش دیگر داشت و در سایر روش‌های دسته‌بندی مشابه دو روش دیگر، عمل می‌کرد. با مقایسه جدول (۵-۱۱) و جدول (۵-۳) این نتیجه حاصل می‌شود که تنها ماشین بردار پشتیبان(خطی) و الگوریتم Bagging (درخت تصمیم) حدود یک درصد کاهش یافته و دقت سایر روش‌های دسته‌بندی به صورت جزئی بهبود یافته است. با مقایسه دو جدول (۵-۴) و (۱۱-۵)، مشاهده می‌شود که دقت دسته‌بندی‌های مختلف در جدول (۵-۱۱)، به صورت آشکاری بهبود یافته است به عنوان مثال الگوریتم Bagging (درخت تصمیم) در جدول (۵-۴) حدود ۱۶٪ بهبود یافته است. در مقایسه با جدول (۵-۷)، جدول (۵-۱۱) به صورت جزئی در بیزین ساده کاهش دقت داشته است و در باقی روش‌های دسته‌بندی دقت افزایش یافته است.

مشابه ویژگی احتمالی، با ترکیب unigram و bigram کاهش یافته، بهره اطلاعات و نرخ بهره ۵۷۰ ویژگی، و مربع کای ۵۶۸ ویژگی استخراج کرد. تابع وزن‌دهی مورد آزمایش در این قسمت PMI است.

جدول (۵-۱۲): ترکیب ویژگی n-gram های کاهش یافته و PMI

	بهره اطلاعات	نرخ بهره	مربع کای
ماشین بردار پشتیبان (چند جمله‌ای)	۷۵/۵۷	۷۵/۵۷	۷۵/۶۶
ماشین بردار پشتیبان (خطی)	۷۴/۷۷	۷۴/۷۷	۷۴/۷۷
بیزین ساده	۶۸/۲۷	۶۸/۲۷	۶۸/۲۷
درخت تصمیم	۷۳/۴۴	۷۳/۴۴	۷۳/۴۴
Bagging (درخت تصمیم)	۷۳/۷۰	۷۳/۷۰	۷۳/۷۰

همانطور که در جدول (۵-۱۲) مشاهده می‌شود، در بهره اطلاعات و نرخ دقت روش‌های دسته‌بندی یکسان است. زیرا در این دو روش عبارت‌ها یکسان هستند. دقت ماشین بردار پشتیبان (چندجمله-ای) در روش مربع کای، به صورت جزئی نسبت به دو روش دیگر بهتر است؛ ولی در سایر روش‌های دسته‌بندی ، مشابه بهره اطلاعات و نرخ بهره است. جدول (۵-۱۲) دقیقا مشابه جدول (۵-۱۱) است.

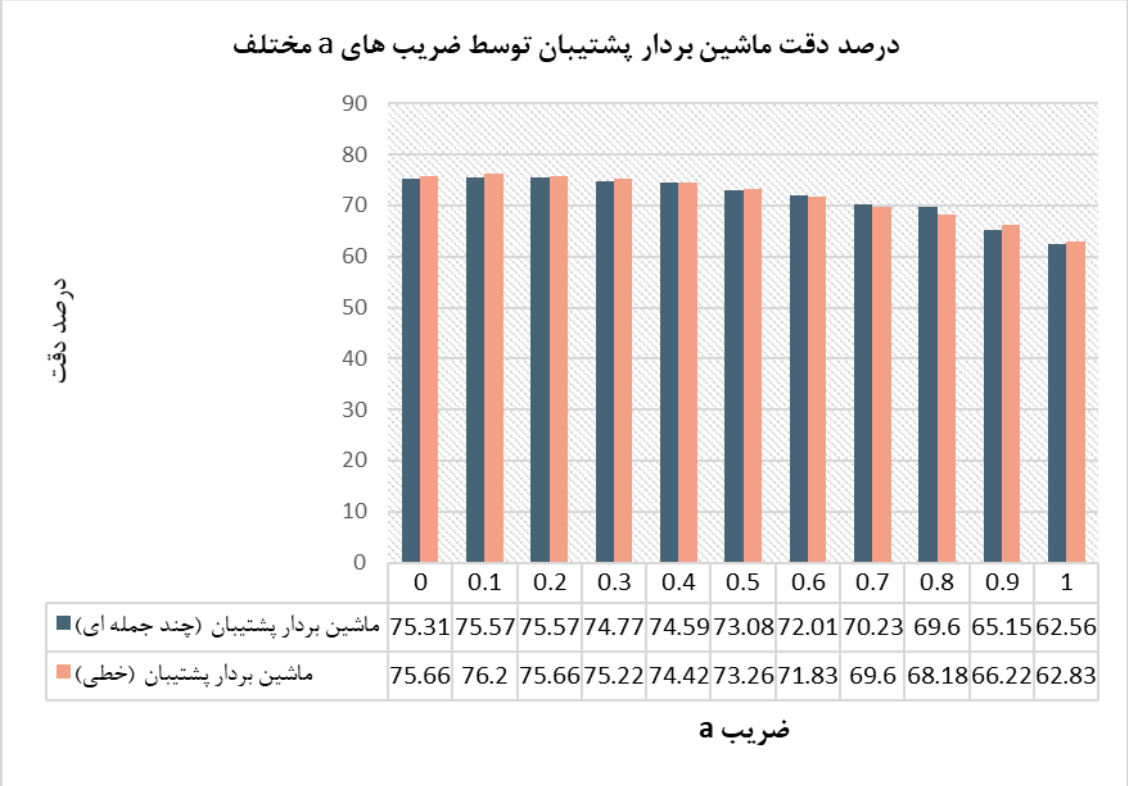
در این بخش عبارت‌های استخراج شده از unigram و bigram با یکدیگر ترکیب شدند. عبارت‌های استخراجی بر اساس ویژگی احتمالی و PMI بدست آمدند. زیرا این دو ویژگی، دقت بالاتری را بدست آوردند. نتایج دسته‌بندی در دو جدول (۵-۱۱) و (۵-۱۲) بررسی شدند. نتایج دسته‌بندی‌های مختلف در دو جدول (۵-۱۱) و (۵-۱۲) مشابه است.

۵-۵- بررسی تاثیر گفته‌های پیشین در دسته‌بندی پاسخ‌ها

در بخش‌های قبل به دسته‌بندی پاسخ‌ها بدون در نظر گرفتن گذشته آن‌ها پرداخته شد. مشاهده شد که با اعمال روش‌های کاهش ویژگی بر روی unigram و bigram، اندازه مجموعه ویژگی، به طرز آشکاری کاهش پیدا می‌کند. این درحالی است که تغییرات دقت در unigram نامحسوس هستند و در مقابل، کاهش دقت با کاهش bigram بسیار آشکار است. با ترکیب ویژگی‌های unigram و bigram، نتایج دقت دسته‌بندی‌ها، نسبت به استفاده از unigram، بصورت جزئی بهبود یافت. این در حالی است که با اعمال روش‌های کاهش ویژگی، اندازه مجموعه ویژگی ترکیب unigram و bigram، ۴ برابر unigram است. به همین دلیل، از unigram برای باقی آزمایش‌ها استفاده می‌شود و از روش بهره اطلاعات برای کاهش ابعاد مجموعه ویژگی‌ها استفاده می‌شود. نوآوری این پژوهش در دسته‌بندی پاسخ‌ها با استفاده از گفتگو پیشین است. همانطور که در فصل چهارم گفته شد، تاثیر سابقه گفتگو به دو صورت وابستگی زمانی و وابستگی ساختاری بررسی می‌شود. با توجه به اینکه یک پاسخ در هر نوبتی می‌تواند وجود داشته باشد. اندازه حافظه ثابت در نظر گرفته نمی‌شود.

۵-۵-۱- وابستگی زمانی

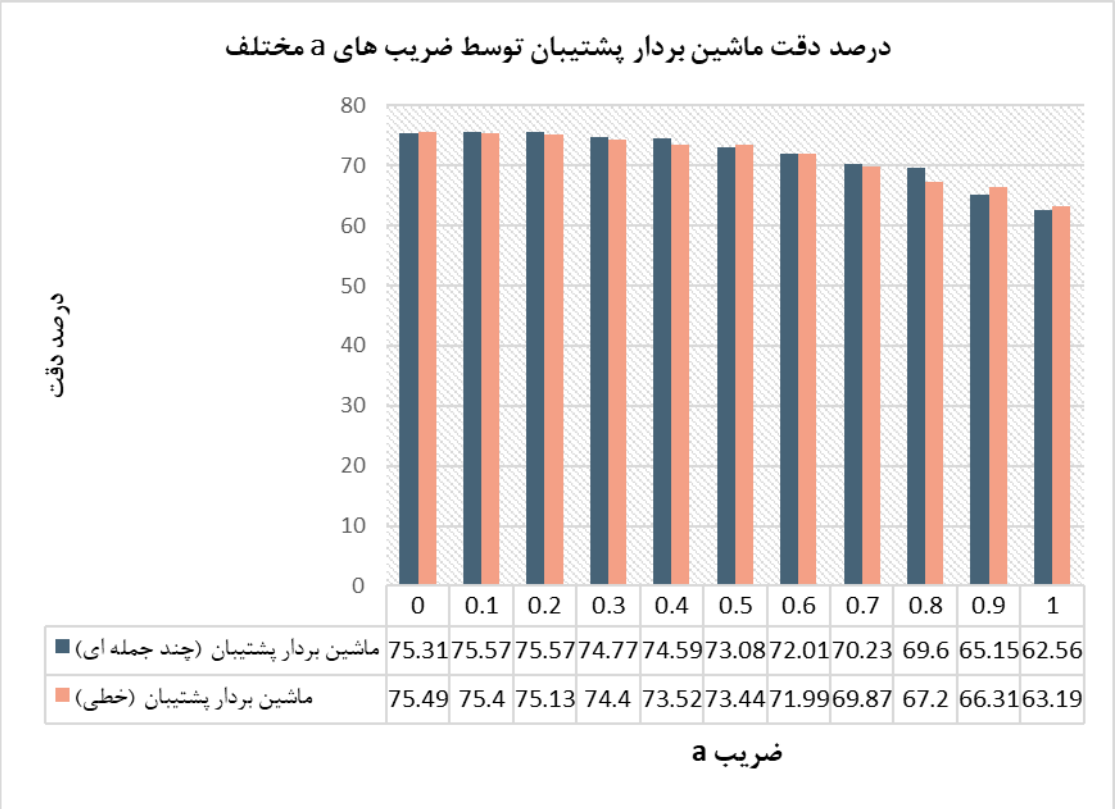
در این مرحله بایستی یک ضریب مناسب انتخاب شود که بوسیله آن بتوان عبارت‌های گفتگوی قبل از پاسخ را در دسته‌بندی تاثیر داد. ضریب مناسب بایستی دقت دسته‌بندی را بهبود دهد. با توجه به اینکه ماشین بردار پشتیبان(خطی و چندجمله‌ای) در تمام دسته‌بندی‌ها دارای درصد دقت بالایی بود، از این روش دسته‌بندی برای انتخاب ضریب مناسب استفاده می‌شود. آزمایش‌ها بر روی unigram و ویژگی احتمالی و اطلاعات متقابل نقطه به نقطه، انجام می‌شود. این مجموعه ویژگی توسط بهره اطلاعات کاهش می‌یابد.



شکل(۵-۱): انتخاب ضریب a مناسب توسط دقت ماشین بردار پشتیبان (چندجمله‌ای و خطی) و ویژگی احتمالی

در شکل (۵-۱) دقت ضریب‌های مختلف a بر روی دقت دسته‌بندی ماشین بردار پشتیبان (هسته خطی و چندجمله‌ای) و ویژگی احتمالی بررسی شده است. ضریب صفر به این معناست که محتوای گفتگوی قبلی در دسته‌بندی تاثیر داده نشده است. در ماشین بردار پشتیبان (چندجمله‌ای) با ضرایب ۰/۱ و ۰/۲ دقت به صورت جزئی افزایش پیدا کرده است و پس آن دقت به صورت نزولی کاهش می-

یابد. در ماشین بردار پشتیبان(خطی) در ضریب ۰/۲ شکستگی وجود دارد. به این صورت که در ضریب ۰/۲ دقت افزایش یافته و پس از آن، با افزایش ضریب a، دقت به صورت نزولی کاهش می‌یابد. با بررسی ضریب‌های مختلف می‌توان نتیجه گرفت که ضریب ۰/۲ برای تاثیر محتوای گفتگوی گذشته مناسب است. ولی این ضریب دقت را به صورت بسیار جزئی افزایش می‌دهد. به طور کلی می‌توان نتیجه گرفت که استفاده از وابستگی زمانی برای تاثیر گفتگوی پیشین، دقت را به صورت مورد انتظار بهبود نمی‌بخشد.



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
ماشین بردار پشتیبان (چند جمله ای)	75.31	75.57	75.57	74.77	74.59	73.08	72.01	70.23	69.6	65.15	62.56
ماشین بردار پشتیبان (خطی)	75.49	75.4	75.13	74.4	73.52	73.44	71.99	69.87	67.2	66.31	63.19

شکل(۵-۲): انتخاب ضریب a مناسب توسط دقت ماشین بردار پشتیبان (چندجملهای و خطی) و PMI

در شکل (۵-۲) دقت ضریب‌های مختلف a بر روی دقت دسته‌بندی ماشین بردار پشتیبان (هسته خطی و چندجمله‌ای) و PMI بررسی شده است. دقت‌های بدست آمده برای ماشین بردار پشتیبان(چندجمله‌ای) در شکل (۵-۲) دقیقا مشابه ماشین بردار پشتیبان(چندجمله‌ای) شکل (۵-۱) است. به طور کلی دقت ماشین بردار پشتیبان(خطی) در شکل (۵-۱) نسبت به ماشین بردار پشتیبان (چندجمله‌ای) بهتر است. با این حال با افزایش ضریب a، دقت ماشین بردار پشتیبان(خطی) کاهش

می‌یابد و در هیچ نقطه‌ای افزایش دقت رخ نمی‌دهد. دقت ماشین بردار پشتیبان (چندجمله‌ای) در ضرایب ۰/۱ و ۰/۲ به صورت جزئی از ۷۵/۳۱٪ به ۷۵/۵۷٪ افزایش می‌یابد و پس از آن با افزایش ضریب a، دقت به صورت نزولی کاهش می‌یابد. با مقایسه شکل‌های (۵-۱) و (۵-۲) به طور کلی این نتیجه حاصل می‌شود که ویژگی احتمالی اندکی بهتر از PMI عمل می‌کند.

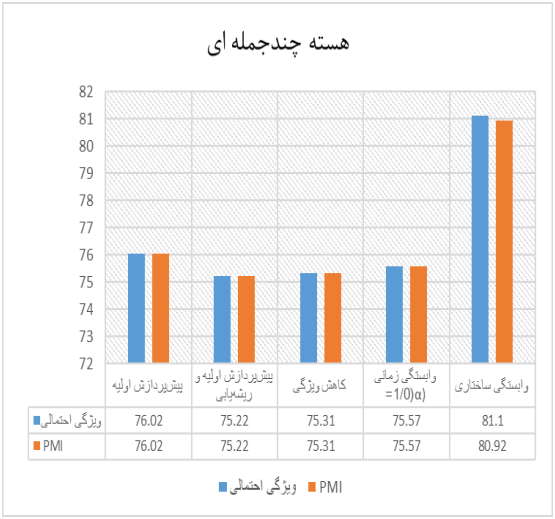
۵-۵-۲- وابستگی ساختاری

همانطور که در فصل چهارم توضیح داده شد، در وابستگی ساختاری، تاثیر سابقه پاسخ براساس نوع گفتگو قبلی انجام می‌شود. این روش برخلاف وابستگی زمانی محتوای گفتگوی قبلی را در دسته‌بندی پاسخ، تأثیر نمی‌دهد. برای این آزمایش unigram انتخاب شده و روش استخراج ویژگی بهره اطلاعات بر روی ویژگی‌های مختلف، اعمال می‌شود.

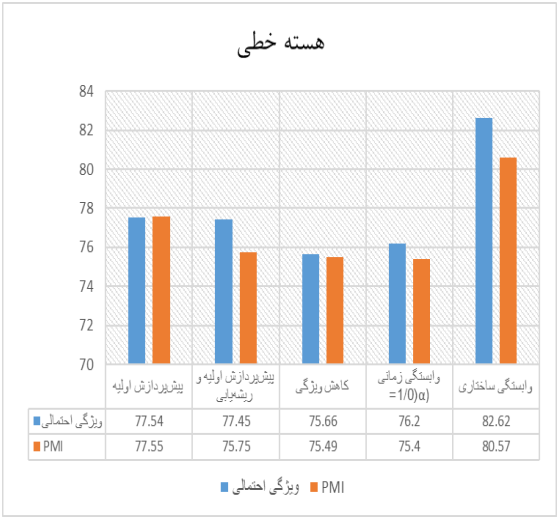
جدول(۵-۱۳): درصد دقت دسته‌بندی‌های مختلف با تاثیرگذاری وابستگی ساختاری

	فرکانس	حضور	TF-IDF	احتمالی	PMI
ماشین بردار پشتیبان(چندجمله‌ای)	۸۱/۱۹	۸۱/۱۹	۸۱/۱۰	۸۱/۱۰	۸۰/۹۲
ماشین بردار پشتیبان(خطی)	۷۱/۵۶	۸۰/۶۵	۶۹/۶۰	۸۲/۶۲	۸۰/۵۷
بیزین ساده	۷۳/۱۷	۷۳/۱۷	۷۳/۱۷	۷۳/۱۷	۷۵/۲۲
درخت تصمیم	۷۶/۰۲	۷۶/۰۲	۷۶/۰۲	۷۶/۰۲	۷۷
الگوریتمBagging(درخت تصمیم)	۷۷/۸۹	۷۷/۸۹	۷۷/۸۹	۷۸/۶۰	۷۸/۵۲

جدول(۵-۱۳) درصد دقت دسته‌بندی‌های مختلف را نشان می‌دهد که در آن سابقه مکالمه براساس وابستگی ساختاری تاثیر گذاشته شده است. همانطور که مشاهده می‌شود، بهبود چشم‌گیری در دقت دسته‌بندی‌ها بوجود آمده است. بهترین دقت توسط ماشین بردار پشتیبان(خطی) و ویژگی احتمالی با مقدار ۸۲/۶۲٪ بدست آمد که در جدول (۵-۷) این مقدار برابر ۷۵/۶۶٪ بود. به عبارت دیگر استفاده از وابستگی ساختاری ۷٪ دقت ماشین بردار پشتیبان(خطی) و ویژگی احتمالی را افزایش داد. با بررسی دقت سایر روش‌های دسته‌بندی در جدول (۵-۱۳)، این نتیجه حاصل می‌شود که روش‌های دیگر نیز به طور متوسط به همین نسبت بهبود یافتند.



الف) ماشین بردار پشتیبان با هسته چند جمله‌ای



ب) ماشین بردار پشتیبان با هسته خطی

شکل (۵-۳): مقایسه دو تابع هسته ماشین بردار پشتیبان

شکل (۵-۳) مقایسه کلی ماشین بردار پشتیبان را نشان می‌دهد. در هر دو نمودار می‌توان بهبود قابل ملاحظه وابستگی ساختاری را مشاهده کرد. به طور کلی هسته خطی دقت بالاتری را نسبت به هسته چندجمله‌ای بدست آورد. در هسته چندجمله‌ای دقت هر دو ویژگی پیشنهادی در تمام قسمت‌ها نزدیک به هم هستند. در مقابل ویژگی احتمالی در هسته خطی، دقت بالاتری را در دو قسمت وابستگی ساختاری و پیش‌پردازش اولیه و ریشه‌یابی بدست آورد.

۵-۶-جمع‌بندی

در این فصل به ارزیابی روش‌های پیشنهادی پرداخته شد. در ابتدا روش مناسب پیش‌پردازش برای اصلاح داده‌ها انتخاب شد. این روش از ترکیب پیش‌پردازش اولیه(حذف علائم نگارشی، یکسان نمودن شکل‌های مختلف حروف و حذف حالت‌های نامتعادل) و ریشه‌یابی بوجود می‌آید. همزمان، دقت دسته‌بندی‌های مختلف را با ویژگی‌هایمختلف بررسی شدند. یک مشکل اساسی در دسته‌بندی‌ها کاهش دقت پایین مربوط به مجموعه ویژگی‌های با ابعاد بزرگ بود. با استفاده از بهره اطلاعات، نرخ بهره و مربع کای ابعاد مجموعه ویژگی کاهش یافت و مشاهده شد هر سه روش، مجموعه ویژگی یکسانی را استخراج کردند. ولی مربع کای در چند ویژگی متفاوت بود. در سایر آزمایش‌ها از بهره اطلاعات برای کاهش ویژگی استفاده گردید. برای تاثیر سابقه پاسخ از دو روش وابستگی زمانی و

وابستگی ساختاری استفاده شد. در وابستگی زمانی بهبود بسیار جزئی بود و یک ضریب مشترک بین تمام روش‌های دسته‌بندی که بتواند به طور موثر دقت دسته‌بندی را بهبود بخشد، وجود نداشت. در مقابل وابستگی ساختاری موجب بهبود تمامی روش‌های دسته‌بندی شد.

فصل ششم:

جمع‌بندی و پیشنهادهایی برای ادامه کار

۶-۱- جمع‌بندی

یکی از بزرگ‌ترین مشکلاتی که در موتورهای جستجوی کنونی وجود دارد، عدم تعامل آن‌ها با کاربر است. این موتورها انبوهی از منابع مرتبط با نیاز کاربر را در اختیار او قرار می‌دهند و کاربر باید اطلاعات مورد نیاز خود را با جستجو در این منابع، پیدا کند. سیستم‌های پرسش و پاسخ، نمونه سیستم‌هایی هستند که اطلاعات مورد نیاز کاربران را بصورت مستقیم فراهم می‌کنند. مهم‌ترین عیب این سیستم‌ها این است که نمی‌توانند با ساختارهای پیچیده و مبهم کار کنند. سیستم‌های پرسش و پاسخ تعاملی، نسخه توسعه یافته سیستم‌های پرسش و پاسخ کلاسیک هستند که از طریق مکالمه به درخواست کاربر پاسخ می‌دهند. استفاده از مکالمه برای تعامل با کاربر مزایای زیادی دارد؛ در صورتی که درخواست کاربر مبهم باشد، سیستم می‌تواند با پرسش سوال‌هایی، این ابهام‌ها را برطرف کند؛ میزان رضایت کاربر را می‌توان از طریق بازخوردهای او محاسبه کرد و کاربر می‌تواند علاوه بر پیدا کردن اطلاعات مورد نیاز خود، درخواست اطلاعات بیشتری کند. این پایان‌نامه بر روی تحلیل پاسخ-های کاربر و سیستم، متمرکز شده است و تلاش می‌کند مفهوم پنهان پاسخ‌ها را در سه دسته مثبت، منفی و خنثی خلاصه‌سازی کند. برای این هدف مراحل زیر در نظر گرفته شده‌اند:

۱- پنج ویژگی فرکانس عبارت، حضور عبارت، TF-IDF، ویژگی احتمالی و اطلاعات متقابل نقطه به نقطه بر روی n-گرم‌های (n=۱) و (n=۲) آزمایش شدند.

۲- انتخاب عبارت‌های مفید در دسته‌بندی قطبی پاسخ‌ها (با استفاده از روش بهره اطلاعات تعداد مناسبی از ویژگی‌ها، انتخاب شدند.)

۳- بررسی روش‌های مختلف یادگیری ماشین با نظارت (ماشین بردار پشتیبان با دو هسته خطی و چند جمله‌ای به عنوان روشی کارآمد برای دسته‌بندی قطبی پاسخ‌ها انتخاب شدند.)

۴- بررسی تاثیر گفتگوی پیشین در دسته‌بندی پاسخ‌ها که به دو صورت وابستگی ساختاری و وابستگی زمانی انجام شد. وابستگی زمانی به صورت بسیار جزئی توانست نتایج را بهبود بخشد. این روش عبارت‌های گفتگوی قبلی را در دسته‌بندی پاسخ فعلی، تاثیر می‌دهد. نحوه تاثیرگذاری این

عبارت‌ها به این صورت است که هر چه فاصله بین پاسخ فعلی با نوبت مورد نظر، کمتر باشد، عبارت-های نوبت مورد نظر ارزش بیشتری در پاسخ فعلی خواهد داشت. در مقابل، در روش وابستگی ساختاری، محتوای نوبت‌های قبلی در پاسخ فعلی تاثیر داده نمی‌شوند. بلکه نوع ارتباط آن‌ها در دسته‌بندی تاثیرگذار است. این روش بهبود قابل توجهی را دسته‌بندی پاسخ‌ها داشت.

۶-۲- پیشنهادهایی برای ادامه کار

یک پاسخ ممکن است حاوی جمله‌های مختلفی باشد. به عنوان مثال ممکن است حاوی سوال باشد و یا اینکه جمله‌ها حاوی قطب‌های متفاوتی باشند. این تحقیق پاسخ را به صورت کلی و گرایش بیشتر آن به یک دسته، کلاس‌بندی می‌کند. یک روش برای بهبود کار این است که بتوان دسته‌بندی را در سطح جمله انجام داد. روش دیگر برای بهبود، این است که خصوصیت‌هایی که کاربر یا سیستم در مورد یک موضوع خاص بیان می‌کنند را تشخیص داد. سپس دسته‌بندی قطبی را بر روی این خصوصیت‌ها اعمال کرد. علاوه بر این موارد می‌توان دسته‌بندی قطبی را به دسته‌های بالاتر (به عنوان مثال دسته‌بندی در پنج دسته) ارتقاء داد.

مراجع

[1] Sharma A. and Dey S. (2012) "A comparative study of feature selection and machine learning techniques for sentiment analysis". in Proceedings of the 2012 ACM research in applied computation symposium. ACM.

[2] Houen S., (2011) "Opinion mining with semantic analysis" **Online:< http://www.diku.dk/forskning/Publikationer/specialer/2011/specialerapport_final_Soren_Houen. pdf>**.

[3] Mukherjee S. and Bhattacharyya P., (2013) "Sentiment analysis: A literature survey" **arXiv preprint arXiv:1304.4520**.

[4] Bouadjenek M. R., Hacid H., and Bouzeghoub M., (2016) "Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms" **Information Systems**. 56: p. 1-18.

[5] Pozzi F. A., Fersini E., Messina E., and Liu B., (2016) "**Sentiment Analysis in Social Networks**". ed., ed. Vol.: Morgan Kaufmann.

[6] Balahur A. and Turchi M. (2012) "Multilingual sentiment analysis using machine translation?". in Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis. Association for Computational Linguistics.

[7] Balahur A. and Turchi M., (2014) "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis" **Computer Speech & Language**. 28(1): p. 56-75.

[8] Tellez E. S., Miranda-Jiménez S., Graff M., Moctezuma D., Suárez R. R., and Siordia O. S., (2017) "A Simple Approach to Multilingual Polarity Classification in Twitter" **Pattern Recognition Letters**.

[9] Pang B., Lee L., and Vaithyanathan S. (2002) "Thumbs up sentiment classification using machine learning techniques". in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics.

[10] Anjaria M. and Guddeti R. M. R. (2014) "Influence factor based opinion mining of Twitter data using supervised learning". in Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on. IEEE.

[11] Ali F., Kwak K.-S., and Kim Y.-G., (2016) "Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification" **Applied Soft Computing**. 47: p. 235-250.

[12] Greengrass E., (2000) "Information retrieval: A survey".

[13] Herzig J., Feigenblat G., Shmueli-Scheuer M., Konopnicki D., Rafaeli A., Altman D., and Spivak D. (2016) "Classifying Emotions in Customer Support Dialogues in Social Media". in SIGDIAL Conference.

[14] Soricut R. and Brill E. (2004) "Automatic Question Answering: Beyond the Factoid". in HLT-NAACL.

[15] Heie M. H., Whittaker E. W., and Furui S., (2012) "Question answering using statistical language modelling" **Computer Speech & Language**. 26(3): p. 193-209.

[16] Adams T., (2017) "AI-Powered Social Bots" **arXiv preprint arXiv:1706.05143**.

[17] Bartl A. and Spanakis G., (2017) "A retrieval-based dialogue system utilizing utterance and context embeddings" **arXiv preprint arXiv:1710.05780**. p.
[18] Liu H., Lin T., Sun H., Lin W., Chang C.-W., Zhong T., and Rudnicky A., (2017) "RubyStar: A Non-Task-Oriented Mixture Model Dialog System" **arXiv preprint arXiv:1711.02781**.
[19] Sacaleanu B., Orasan C., Spurk C., Ou S., Ferrandez O., Kouylekov M., and Negri M. (2008) "Entailment-based question answering for structured data". in 22nd International Conference on on Computational Linguistics: Demonstration Papers. Association for Computational Linguistics.
[20] Saif H., Fernández M., He Y., and Alani H., (2014) "On stopwords, filtering and data sparsity for sentiment analysis of twitter".
[21] Raschka S., (2014) "Naive bayes and text classification i-introduction and theory" **arXiv preprint arXiv:1410.5329**.
[22] Paltoglou G. and Thelwall M. (2010) "A study of information retrieval weighting schemes for sentiment analysis". in Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics.
[23] Tripathy A., Agrawal A., and Rath S. K., (2016) "Classification of sentiment reviews using n-gram machine learning approach" **Expert Systems with Applications**. 57: p. 117-126.
[24] Bespalov D., Bai B., Qi Y., and Shokoufandeh A. (2011) "Sentiment classification based on supervised latent n-gram analysis". in Proceedings of the 20th ACM international conference on Information and knowledge management. ACM.
[25] Gimpel K., Schneider N., O'Connor B., Das D., Mills D., Eisenstein J., Heilman M., Yogatama D., Flanigan J., and Smith N. A. (2011) "Part-of-speech tagging for twitter: Annotation, features, and experiments". in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics.
[26] Kamal A. and Abulaish M. (2013) "Statistical features identification for sentiment analysis using machine learning techniques". in Computational and Business Intelligence (ISCBI), 2013 International Symposium on. IEEE.
[27] Church K. W. and Hanks P., (1989) "Word association norms, mutual information, and lexicography, Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics".
[28] Turney P. D. (2002) "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews". in Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics.
[29] Karegowda A. G., Manjunath A., and Jayaram M., (2010) "Comparative study of attribute selection using gain ratio and correlation based feature selection" **International Journal of Information Technology and Knowledge Management**. 2(2): p. 271-277.
[30] Li Y., Xie M., and Goh T., (2009) "A study of mutual information based feature selection for case based reasoning in software cost estimation" **Expert Systems with Applications**. 36(3): p. 5921-5931.
[31] Bouma G., (2009) "Normalized (pointwise) mutual information in collocation extraction" **Proceedings of GSCL**. p. 31-40.

[32] Tan S. and Zhang J., (2008) "An empirical study of sentiment analysis for chinese documents" **Expert Systems with applications**. 34(4): p. 2622-2629.
[33] Vinodhini G. and Chandrasekaran R., (2012) "Sentiment analysis and opinion mining: a survey" **International Journal**. 2(6): p. 282-292.
[34] Fu G. and Wang X. (2010) "Chinese sentence-level sentiment classification based on fuzzy sets". in Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics.
[35] Madhoushi Z., Hamdan A. R., and Zainudin S. (2015) "Sentiment analysis techniques in recent works". in Science and Information Conference (SAI), 2015. IEEE.
[36] Kalaivani P. and Shunmuganathan K., (2013) "Sentiment classification of movie reviews by supervised machine learning approaches" **Indian Journal of Computer Science and Engineering**. 4(4): p. 285-292.
[37] Gonçalves P., Araújo M., Benevenuto F., and Cha M. (2013) "Comparing and combining sentiment analysis methods". in Proceedings of the first ACM conference on Online social networks. ACM.
[38] Mohammad S. M. and Turney P. D., (2013) "Crowdsourcing a word–emotion association lexicon" **Computational Intelligence**. 29(3): p. 436-465.
[39] Miller G. A., (1995) "WordNet: a lexical database for English" **Communications of the ACM**. 38(11): p. 39-41.
[40] Baccianella S., Esuli A., and Sebastiani F. (2010) "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". in LREC.
[41] Zhang A. X., Culbertson B., and Paritosh P., (2017) "Characterizing Online Discussion Using Coarse Discourse Sequences".
[42] Kendrick K. H., (2017) "Using Conversation Analysis in the Lab" in, Editor and Editors, Taylor & Francis.
[43] Liu C., Ishi C., and Ishiguro H., (2017) "Turn-taking Estimation Model Based on Joint Embedding of Lexical and Prosodic Contents" **Proc. Interspeech 2017**. p. 1686-1690.
[44] Heldner M. and Edlund J., (2010) "Pauses, gaps and overlaps in conversations" **Journal of Phonetics**. 38(4): p. 555-568.
[45] Murray G., Carenini G., and Ng R. (2010) "Generating and validating abstracts of meeting conversations: a user study". in Proceedings of the 6th International Natural Language Generation Conference. Association for Computational Linguistics.
[46] Maryam T., (2013) "Dialogue act recognition in synchronous and asynchronous conversations," University of British Columbia.
[47] ZIĘBA M., TOMCZAK J. M., and BRZOSTOWSKI K., "ASKING RIGHT QUESTIONS BASING ON INCOMPLETE DATA USING RESTRICTED BOLTZMANN MACHINES" **Information Systems Architecture and Technology**. p. 23.
[48] Mayfield E., Adamson D., and Rosé C. P. (2012) "Hierarchical conversation structure prediction in multi-party chat". in Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics.
[49] McTear M., Callejas Z., and Griol D., (2016) "Dialog Management" in The Conversational Interface, Springer. p. 209-233.

[50] Griol D., Callejas Z., López-Cózar R., and Riccardi G., (2014) "A domain-independent statistical methodology for dialog management in spoken dialog systems" **Computer Speech & Language**. 28(3): p. 743-768.
[51] Asri L. E., Schulz H., Sharma S., Zumer J., Harris J., Fine E., Mehrotra R., and Suleman K., (2017) "Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems" **arXiv preprint arXiv:1704.00057**.
[52] Griol D., Hurtado L. F., Segarra E., and Sanchis E., (2008) "A statistical approach to spoken dialog systems design and evaluation" **Speech Communication**. 50(8): p. 666-682.
[53] Konstantinova N. and Orasan C., (2012) "Interactive question answering" **Emerging Applications of Natural Language Processing: Concepts and New Research**. p. 149-169.
[54] Arora S., Batra K., and Singh S., (2013) "Dialogue System: A Brief Review" **arXiv preprint arXiv:1306.4134**.
[55] Ravichandran D. and Hovy E. (2002) "Learning surface text patterns for a question answering system". in Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics.
[56] Riloff E. and Thelen M. (2000) "A rule-based question answering system for reading comprehension tests". in Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems-Volume 6. Association for Computational Linguistics.
[57] Chali Y., Hasan S. A., and Mojahid M., (2015) "A reinforcement learning formulation to the complex question answering problem" **Information Processing & Management**. 51(3): p. 252-272.
[58] Kelly D., Wacholder N., Rittman R., Sun Y., Kantor P., Small S., and Strzalkowski T., (2007) "Using interview data to identify evaluation criteria for interactive, analytical question-answering systems" **Journal of the Association for Information Science and Technology**. 58(7): p. 1032-1043.
[59] Oraby S., Gundecha P., Mahmud J., Bhuiyan M., and Akkiraju R. (2017) "How May I Help You?: Modeling Twitter Customer ServiceConversations Using Fine-Grained Dialogue Acts". in Proceedings of the 22nd International Conference on Intelligent User Interfaces. ACM.
[60] Harabagiu S., Hickl A., Lehmann J., and Moldovan D. (2005) "Experiments with interactive question-answering". in Proceedings of the 43rd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics.
[61] Joachims T., "Text Categorization with Support Vector Machines: Learning with Many Relevant".

Abstract

Contemporary search engines are not optimized to respond to users' exact inquiries. These engines simply list some relevant resources for users. Providing information through interaction is the main idea of a questions and answers (Q&A) system. Most researches which conducted on these systems focuses on the processing of questions. The processing of the answer is usually limited to choosing the appropriate answer to the previous question. In this research, polarity classification is applied to the responses of an interactive questions and answers system. Polar classification method is one of the most well-known branches in text processing. This method categorizes inputs into three positive, negative and neutral categories according to their internal content. Choosing the right features for classification is of course crucial.

In this study, two new features called probability characteristics and developed point-wise mutual information have been introduced. These features, considering their tendency to one of three positive, negative and neutral categories, will set the terms. In addition to these two features, three popular features have been used in polarization are also tested. These tests are performed on unigrams and bigrams. To reduce the size of the data, information gain(GI), gain ratio (GR) and CHI square are utilized.

Hitherto, polarity classification method has not been utilized on interactive texts. In a conversation, previous utterance can influence on the polarity of the answer. In this study, the effect of the previous statements has been investigated in terms of both time dependence and structural dependence. The dependence of time affects the content of the preceding statements in categorization, while structural dependence focuses on its type. The best possible precision with the unigram-based feature is 77.5%. By applying the attrition, the precision becomes 75.66% and its structural dependence reaches 82.62%

Keywords: Polarity classification, Interactive question answering systems, Conversation history

Shahrood University of Technology

Faculty of Computer Engineering
M.Sc. Thesis in Artificial Intelligence Engineering

Answer summarization of an automatic dialogue into limited set using artificial intelligence technique

By:
Moloud Ayat

Supervisor:
Dr. Morteza Zahedi

Jan 2018