



بسم الله الرحمن الرحيم



دانشگاه اصفهان

دانشکده فنی - مهندسی

گروه مهندسی برق

## پایان نامه کارشناسی رشته مهندسی برق گرایش مخابرات

<عنوان پایان نامه: پردازش زبان طبیعی در فارسی>

استاد راهنما:

دکتر فرزاد پرورش

پژوهشگر:

سید امیر موسوی مبارکه

شهریور ۱۳۹۸

اکنون که با عنایت و یاری خداوند متعال مراحل پژوهش، تدوین و نگارش تحقیق به پایان رسیده است؛ بر خود واجب میدانم از عزیزانی که طی مراحل مختلف از راهنمایی و یاری آنها بهره بردم سپاسگزاری نمایم.

در ابتدا بر خود واجب می دانم که از پدر و مادرم نهایت تشکر و قدردانی را به عمل آورم که در تمام مراحل زندگی مرا یاری نموده اند .

استاد ارجمند جناب دکتر پرورش به پاس نعمت توفیق و ادای وظیفه بر خود لازم می دانم که صمیمانه ترین قدردانی های خویش را نثار شما بنمایم که دلسوزانه مجموعه علم و دانش خویش را در اختیار من گذاشتید و در طی این مسیر چهارساله با وجود دل سردی و ناامیدی ها همیشه فضای پر انرژی را برای دانشجویان ایجاد نموده اید .

در پایان لازم می دانم از تمامی افرادی که بطور مستقیم و غیر مستقیم در به ثمر نشستن این پژوهش مرا مرهون مساعدت و همکاری خود نمودند؛ صمیمانه تشکر و قدردانی کنم.

## چکیده

هدف اصلی تحقیق، بررسی پردازش زبان طبیعی (NLP: natural language processing) می باشد. پردازش زبان طبیعی یکی از زیر شاخه های پر اهمیت در حوزه ی علوم رایانه و هوش مصنوعی است؛ که به تعامل بین کامپیوتر و زبان های طبیعی انسان می پردازد؛ به بیان دیگر پردازش زبان طبیعی قادر ساختن ماشین ها به درک گفتار یا نوشتار تولید شده در قالب یک زبان طبیعی می باشد. ظهور پردازش زبان طبیعی به دهه ۱۹۵۰ میلادی باز می گردد زمانی که آلن تورینگ مقاله ای به نام آزمون تورینگ (Turing Test) را بیان کرد. آزمون تورینگ معیاری بود برای بررسی میزان هوشمندی ماشین ها در حقیقت آزمون تورینگ تستی از توانایی ماشین است (در ادامه مفصل مورد بحث قرار می گیرد). پردازش زبان طبیعی در تعدادی از زبان های اصلی دنیا از جمله انگلیسی محقق شده است در این تحقیق قرار است که پردازش زبان طبیعی در زبان فارسی مورد بررسی قرار بگیرد. در این مسیر سعی شده است فرمول ها و نکات آماری مورد نیاز به صورت کلی بیان شود و از بررسی فرمول ها به صورت تخصصی اجتناب شود. در قسمت برنامه نویسی از زبان python استفاده شده است و در هر قسمت توضیح مختصری در مورد کد نوشته شده داده می شود. کتابخانه های متعددی برای پردازش زبان طبیعی در پایتون وجود دارد که کتابخانه های `numpy`, `pandas`, `nltk`, `spacy`, `skitlearn`, `metrix` از مهم ترین آن ها می باشد. دیتا مورد استفاده داده های سایت دیجیکالا می باشد که شامل حدود ۱۵ هزار نظر کاربر می باشد. برای پردازش از دو روش استفاده شده است روش اول طبقه بندی متن (Text Classification) و روش دوم تحلیل احساساتی (Sentiment Analysis) می باشد. مرحله ی بعدی پردازش مدل سازی مبحث (Topic Modeling) می باشد. هر سه مدل به صورت کامل در فصل های جداگانه توضیح داده خواهد شد. در آخر نظر جدیدی به مدل ها داده می شود و خروجی آن بررسی رضایت مشتری از کالای خریداری شده، نقطه مثبت کالا در صورت رضایت مشتری و نقطه ضعف کالا در صورت نارضایتی مشتری می باشد.

## فهرست مطالب

عنوان	صفحه
فصل اول مقدمه .....	۱
فصل سوم هوش مصنوعی .....	۳
۱-۲- تاریخچه .....	۳
۲-۲- آزمون تورینگ .....	۷
۳-۲- آیا این هوشمند سازی است؟ .....	۹
در بیش از شصت سالی که آزمون تورینگ در حوزه هوش مصنوعی حضور داشته است، انتقادات مختلفی به آن وارد شده که بخش بزرگی از آن‌ها بر این موضوع استوار بوده‌اند که آیا این آزمون معیار خوبی برای تشخیص هوشمندی یک سیستم است؟ .....	
۴-۲- مراحل هوشمند سازی .....	۱۱
در این قسمت مراحل هوش مصنوعی می‌تواند به آن دست یابد را بررسی می‌کنیم: .....	
۵-۲- انواع هوش مصنوعی : .....	۱۲
۶-۲- شاخه های هوش مصنوعی : .....	۱۵
فصل سوم پردازش زبان طبیعی .....	۲۹
۱-۳- پردازش زبان طبیعی .....	۲۹
فصل چهارم طبقه بندی متن .....	۳۴
۱-۴- طبقه بندی متن (TEXT CLASSIFICATION) .....	۳۴
۲-۴- مروری کلی بر مباحث پایه ای یادگیری ماشین .....	۳۵
۳-۴- مراحل یادگیری ماشین با نظارت (UNSUPERVISED LEARNING) .....	۳۶
۴-۴- معیارهای طبقه بندی (CLASSIFICATION METRICS) .....	۴۱
۵-۴- ماتریس پیریشانی (CONFUSION MATRIX) .....	۴۲
۶-۴- استخراج ویژگی ها (FEATURE EXTRACTION) .....	۴۷
۷-۴- آشنایی با SKIT-LEARN در PYTHON و اعمال آن بر داده های واقعی .....	۵۲
فصل پنجم تحلیل احساسات .....	۵۵

۵۵.....	۱-۵- تحلیل احساسات (SENTIMENT ANALYSIS)
	۲-۵- VALANCE AWARE DICTIONARY FOR SENTIMENT
۵۶.....	(VADER)REASONING
۵۸.....	۳-۵- کدنویسی تحلیل احساسات در پایتون:
۶۲.....	فصل شیشم مدل سازی موضوع
۶۲.....	۱-۶- مدل سازی موضوع (TOPIC MODELING)
۶۳.....	۲-۶- (LDA) LATENT DIRICHLET ALLOCATION
۶۴.....	۳-۶- کدنویسی مدل سازی مبحث:
۶۷.....	۴-۶- نتیجه و جمع بندی نهایی:
۶۸.....	مراجع
۶۹.....	ضمیمه ۱
۷۱.....	ضمیمه ۲
۷۲.....	ضمیمه ۳

## فهرست شکل‌ها

عنوان	صفحه
شکل ۱ هوش مصنوعی واکنش گر .....	۱۳
شکل ۲ یادگیری عمیق .....	۲۱
شکل ۳ مقایسه یادگیری ماشین و یادگیری عمیق .....	۲۲
شکل ۴ یادگیری عمیق .....	۲۴
شکل ۵ مراحل یادگیری با نظارت .....	۳۶
شکل ۶ اکتساب داده ها .....	۳۶
شکل ۷ تمیز کردن داده ها .....	۳۷
شکل ۸ تقسیم داده ها .....	۳۸
شکل ۹ آموزش مدل .....	۳۸
شکل ۱۰ تست کردن مدل .....	۳۹
شکل ۱۱ بهبود مدل .....	۴۰
شکل ۱۲ گسترش مدل .....	۴۰
شکل ۱۳ ماتریس پیریشانی .....	۴۳
شکل ۱۴ مثال ماتریس پیریشانی .....	۴۶
شکل ۱۷ خروجی طبقه بندی متن .....	۵۳
شکل ۱۸ خروجی تحلیل احساسات .....	۶۱
شکل ۱۹ خروجی مدل سازی مبحث .....	۶۶



## فهرست جدول‌ها

صفحه	عنوان
۴۸.....	جدول ۱ Count Vectorization
۵۰.....	جدول ۲ TF-IDF

## فصل اول

### مقدمه

این روزها همه جا صحبت از هوش مصنوعی (Artificial Intelligence) است که به طور مخفف با عنوان AI از آن یاد می‌شود. در این مبحث، هوشمندی ماشین‌ها، در قیاس با هوش طبیعی موجود در انسان‌ها مورد بررسی قرار می‌گیرد. نقش هوش مصنوعی هر روز در زندگی ما بیشتر می‌شود. شروع توسعه‌ی این تکنولوژی در واقع به خیلی قبل‌تر برمی‌گردد؛ یعنی زمانی در دهه‌ی ۵۰ میلادی که دانشگاه دارتموث (Dartmouth College) در ایالات متحده یک پروژه‌ی تحقیقات تابستانی را به هوش مصنوعی اختصاص داد. ریشه‌های هوش مصنوعی را حتی می‌توان در عمق بیشتری از تاریخ و در فعالیت‌های آلن نیوئل (Allen Newell)، هربرت ای. سیمون (Herbert A. Simon) و آلن تورینگ (Alan Turing) جست‌وجو کرد. اصطلاح هوش مصنوعی برای اولین بار توسط جان مکارتی (که از آن به‌عنوان پدر علم و دانش تولید ماشین‌های هوشمند یاد می‌شود) استفاده شد ولی بزرگترین جهش را می‌توان در آزمون مشهور تورینگ در سال ۱۹۵۰ مشاهده کرد. این مقاله یکی از اولین اسنادی است که در آن به وجود آمدن ماشین‌های هوشمند پیش‌بینی شده است.

اهمیت گسترش هوش مصنوعی بر کسی پوشیده نیست حتی باعث نگرانی عده‌ای از افراد و احساس خطر شده است. در مقابل افرادی بر این معتقدند که هوش مصنوعی باعث سهولت در زندگی انسان‌ها می‌شود و مسیری است کاملاً سازنده.

یکی از زیر شاخه‌های اصلی هوش مصنوعی پردازش زبان طبیعی می‌باشد در حقیقت به صورت کلی پردازش زبان طبیعی به دنبال راهی برای ایجاد ارتباط بین زبان انسان و ماشین می‌باشد که بدون شک محقق شدن این موضوع قدمی بزرگ بر هوشمند سازی ماشین‌ها می‌باشد.

در فصل اول مقاله نگاهی کوتاه به تاریخچه ی ظهور هوش مصنوعی و زیر شاخه های آن انداخته می شود و جایگاه پردازش زبان طبیعی را در هوش مصنوعی مورد بررسی قرار می دهیم و نگاه کلی بر آزمون تورینگ می اندازیم و اهمیت آن در سنجش هوش مصنوعی مورد بحث قرار می دهیم. امروزه انسان ها سعی می کنند بیشتر کارهای روزمره ی خود را در راستای کاهش وقت و کاهش هزینه رفت و آمد از طریق شبکه های مجازی انجام دهند به خصوص خرید کالاهای غالباً غیر ضروری. از وب سایت های در دسترس می توان به آمازون اشاره کرد که در ابعاد جهانی اقدام به فروش انواع کالا کرده است متأسفانه در ایران با توجه به تحریم ، سایت آمازون از دسترس کاربران ایرانی خارج می باشد که منجر به شکل گیری وب سایت های داخلی جهت فروش کالا شده است که از جمله می توان به وب سایت دیجیکالا اشاره کرد .

این وب سایت ها قابلیت هایی از جمله دیدن نظرات کاربرانی که کالای مورد نظر را خریده اند ، میزان رضایت خریداران از کالای مورد نظر و نقاط ضعف و قوت کالا می توان اشاره کرد که این قابلیت ها باعث می شود که کاربران در خرید هر کالایی با دید بازتری تصمیم بگیرند. همزمان شدن این موضوع با ظهور هوش مصنوعی توجه کد نویسان و تحلیل گران را جلب کرده است و نتیجه ی آن ادغام کردن این دو موضوع شده است در واقع تحلیل گران به دنبال ماشینی هوشمند بودند که بتواند داده های بزرگ را تحلیل کند، به موازات آن مهندسان در راستای هوشمند سازی کامپیوتر ها به دنبال روشی بودن که کامپیوتر ها را قادر به درک زبان انسان ها نمایند. نتیجه این تلاش ظهور پردازش زبان طبیعی شد که در فصل دوم به صورت مختصر توضیح داده خواهد شد.

داده هایی که در این مقاله مورد بررسی قرار گرفته است تحلیل داده های سایت دیجیکالا می باشد که شامل ۱۵ هزار نظر کاربر در مورد کالا های مختلف می باشد. روش های متعددی برای پردازش زبان طبیعی بیان شده است دو روش قابل توجه روش طبقه بندی متن (Text Classification) و روش دوم تحلیل احساساتی (Sentiment Analysis) می باشد. که هر کدام از روش ها به صورت جداگانه در فصل سوم و چهارم مورد بررسی قرار می گیرد. در هر دو روش بیان شده مدلی ایجاد می شود که قابلیت دریافت نظر جدیدی می باشد که خروجی آن بررسی مثبت بودن نظر یا منفی بودن نظر می باشد.

یکی از موضوعات جذاب در پردازش زبان طبیعی مدل سازی مبحث (Topic Modeling) می باشد در این مرحله از پردازش در این موضوع خاص مدلی ایجاد می شود که خروجی آن مشخص کردن نکته ی مثبت کالا در صورتی که کاربر خرید کالا را پیشنهاد می کند و نکته ی منفی کالا در صورتی که کاربر خرید کالا را پیشنهاد نمی کند. که این موضوع در فصل پنجم به صورت کامل بیان می شود. در هر فصل علاوه بر توضیحات گفته در هر قسمت کد نویسی انجام شده به زبان پایتون اضافه شده است و توضیحات لازم در مورد کد داده شده است. توجه شود در این مقاله پایتون آموزش داده نمی شود ولی در هر قسمت کد توضیحات لازم مربوط به آن کد داده شده است.

## فصل سوم

### هوش مصنوعی

#### ۱-۲- تاریخچه

هوش مصنوعی برای نخستین بار توسط جان مکاریتی (John Mccorthy) که از آن به عنوان پدر علم و دانش تولید ماشین های هوشمند یاد می شود، استفاده شد. آقای جان مکاریتی مخترع یکی از زبان های برنامه نویسی هوش مصنوعی به نام lisp نیز است. با این عنوان می توان به هویت هوشمند یک ابزار مصنوعی اشاره کرد. پروفیسور جان مک کارتی در سال ۱۹۲۷ در شهر بوستون متولد شد. وی درجه کارشناسی ارشد خود را در رشته ریاضی در سال ۱۹۴۸ از انستیتو کالیفرنیا و مدرک دکترای خود را از دانشگاه پرینستون در سال ۱۹۵۱ دریافت کرد. او با ادامه تحصیل در رشته علوم رایانه موفق به دریافت درجه استادی در این رشته، از دانشگاه استنفورد شد و از سال ۱۹۶۵ تا ۱۹۸۰ سرپرستی آزمایشگاه هوش مصنوعی دانشگاه استنفورد را برعهده داشت. پیش از بوجود آمدن علوم الکترونیک، هوش مصنوعی توسط فلاسفه و ریاضی دانانی نظیر بول که اقدام به ارائه قوانین و نظریه هایی در باب منطق کردند، مطرح شده بود. با اختراع رایانه های الکترونیک در سال ۱۹۴۳، هوش مصنوعی دانشمندان را به چالشی بزرگ فراخواند. در بادی امر، چنین به نظر می رسید که این فن آوری در نهایت قادر به شبیه سازی رفتارهای هوشمندانه خواهد بود. با وجود مخالفت گروهی از متفکرین با هوش مصنوعی که با دیده تردید به کارآمدی آن می نگریستند

فقط پس از چهار دهه، شاهد تولد ماشین‌های شطرنج باز و دیگر سامانه‌های هوشمند در صنایع گوناگون هستیم.

نقطه آغاز علم هوش مصنوعی را می‌توان به بعد از جنگ جهانی دوم نسبت داد، در آن زمان واینر با توجه به مسائل سایبرنتیک زمینه را برای پیشرفت هوش مصنوعی به وجود آورد و در سال ۱۹۵۰ تورینگ آزمایشی را برای اثبات هوشمند بودن یک ماشین پیشنهاد داد سپس در سال ۱۹۵۶ گروهی از علاقه‌مندان به هوش مصنوعی در کالج دارتموت گرد هم آمدند و پژوهش‌های وسیعی را برای هوش مصنوعی آغاز کردند.

اما با این حال هوش مصنوعی علمی است بسیار جوان و روبه رشد. شروع هوش مصنوعی به معنای واقعی به سال ۱۹۵۰ باز می‌گردد یعنی زمانی که آلن تورینگ مقاله خود را درباره ساخت ماشین هوشمند به رشته تحریر درآورد. در این مقاله تورینگ روشی را برای تشخیص هوشمندی ماشین‌ها پیشنهاد داد که در ادامه تعریفی از هوش مصنوعی بیان می‌شود و در قسمت بعدی به صورت کامل در مورد آزمون تورینگ بحث می‌شود.

## ۲-۱- تعریف هوش مصنوعی

هنگامی که سخن از هوش مصنوعی به میان می‌آید، هرکس به ظن خود تعریفی از آن ارائه می‌کند. این مبحث برای برخی به پیچیدگی ساخت یک مغز شبیه‌سازی شده و برای برخی دیگر به سادگی یک آدم آهنی است. این تفاوت رویکردها نسبت به یک مفهوم واحد، نه صرفاً ناشی از عدم آگاهی عموم از آن، که چه بسا نشأت گرفته از اختلاف نظرهای متعددی است که دانشمندان این حوزه با یکدیگر دارند. این اختلاف نظرها از یک‌سو و وجود لغات و اصطلاحات متعدد در حوزه هوش مصنوعی از سوی دیگر موجب شده تا درک آن بسیار پیچیده به نظر بیاید. در مطلب پیش رو، به برخی از تعاریف و اصطلاحات مورد استفاده در این زمینه پرداخته شده است.

تعریف هوش مصنوعی آن را به عنوان شاخه‌ای از علوم کامپیوتر مشخص می‌کند که با خودکارسازی رفتارهای هوشمندانه سروکار دارد. بخش سخت ماجرا این است: از آنجا که خود هوش را نمی‌توانیم به درستی تعریف کنیم، امکان تعریف دقیق هوش مصنوعی هم وجود ندارد. به طور کلی اصطلاح هوش مصنوعی برای تشریح کردن سیستم‌هایی به کار می‌رود که هدف آن‌ها استفاده از ماشین‌ها برای تقلید و شبیه‌سازی هوش انسانی و رفتارهای مرتبط با آن است. این هدف گاه ممکن است با استفاده از الگوریتم‌های ساده و الگوهای از پیش تعیین شده محقق شود، ولی گاهی هم نیاز به الگوریتم‌ها فوق‌العاده پیچیده دارد.

هوش مصنوعی، هوش صناعی یا هوش ماشینی (Artificial Intelligence) هوشی که یک ماشین در شرایط مختلف از خود نشان می‌دهد، گفته می‌شود. به عبارت دیگر هوش مصنوعی به سیستم‌هایی گفته می‌شود که می‌توانند واکنش‌هایی مشابه رفتارهای هوشمند انسانی از جمله درک شرایط پیچیده، شبیه‌سازی فرایندهای تفکری و شیوه‌های استدلالی انسانی و پاسخ موفق به آنها، یادگیری و توانایی کسب دانش و استدلال برای حل مسایل را داشته باشند. بیشتر نوشته‌ها و مقاله‌های مربوط به هوش مصنوعی، آن را به عنوان (دانش شناخت و طراحی عامل‌های هوشمند) تعریف کرده‌اند.

هنوز تعریف دقیقی برای هوش مصنوعی که مورد توافق دانشمندان این علم باشد ارائه نشده است و این به هیچ وجه مایه تعجب نیست. چرا که مقوله مادر و اساسی‌تر از آن، یعنی خود هوش هم هنوز به‌طور همه‌جانبه و فراگیر تن به تعریف نداده است. در واقع می‌توان نسل‌هایی از دانشمندان را سراغ گرفت که تمام دوران زندگی خود را صرف مطالعه و تلاش در راه یافتن جوابی به این سؤال عمده نموده‌اند که: هوش چیست؟

اما اکثر تعریف‌هایی که در این زمینه ارائه شده‌اند بر پایه یکی از ۴ باور زیر قرار می‌گیرند:

سیستم‌هایی که به‌طور منطقی فکر می‌کنند

سیستم‌هایی که به‌طور منطقی عمل می‌کنند

سیستم‌هایی که مانند انسان فکر می‌کنند

سیستم‌هایی که مانند انسان عمل می‌کنند

اینکه هوش مصنوعی چیست و چه تعریفی می‌توان از آن بیان نمود؟ مبحثی است که تاکنون دانشمندان به یک تعریف جامع در آن نرسیده‌اند و هریک تعریفی را ارائه نموده‌اند که در زیر نمونه‌ای از این تعاریف آمده‌است.

هنر ایجاد ماشین‌هایی که وظایفی را انجام می‌دهند که انجام آن‌ها توسط انسان‌ها نیاز به هوش دارد (کورزویل - ۱۹۹۰)

مطالعه اینکه چگونه کامپیوترها را قادر به انجام اعمالی کنیم که در حال حاضر، انسان آن اعمال را بهتر انجام می‌دهد. (ریچ و نایت - ۱۹۹۱)

خودکارسازی فعالیت‌هایی که ما آن‌ها را به تفکر انسانی نسبت می‌دهیم. فعالیت‌هایی مثل تصمیم‌گیری، حل مسئله، یادگیری و ... (بلمن - ۱۹۷۸)

تلاشی نو و مهیج برای اینکه کامپیوترها را قادر به فکر کردن کنیم. ماشین‌هایی با فکر و حس تشخیص واقعی (هاگلند - ۱۹۸۵)

یک زمینه تخصصی که به دنبال توضیح و شبیه‌سازی رفتار هوشمندانه به وسیله فرایندهای کامپیوتری است. (شالکوف - ۱۹۹۰)

مطالعه محاسباتی که درک، استدلال و عمل کردن را توسط ماشین‌ها را ممکن می‌سازد. (وینستون - ۱۹۹۲)

توانایی دست یافتن به کارایی در حد انسان در همه امور شناختی توسط رایانه (آلن تورینگ - ۱۹۵۰)

هوش مصنوعی دانش و مهندسی ساخت ماشین‌های هوشمند و به خصوص برنامه‌های رایانه‌ای هوشمند است. هوش مصنوعی با وظیفه مشابه استفاده از کامپیوترها برای فهم چگونگی هوش انسان مرتبط است، اما مجبور نیست خودش را به روش‌هایی محدود کند که بیولوژیکی باشند. (جان مک‌کارتی - ۱۹۸۰)

تمام تعریف‌های بیان شده در قسمت بالا دیدگاه‌های مختلف به موضوع هوش مصنوعی بود اولین تعریف کامل و قابل لمسی از هوش مصنوعی توسط تورینگ انجام شد که به آزمون تورینگ معروف می‌باشد.

## ۲-۲- آزمون تورینگ

در سال ۱۹۵۰ آلن تورینگ در مقاله‌ای با عنوان «ساز و کار رایانش و هوشمندی» برای نخستین‌بار آزمون تورینگ را به جهانیان معرفی کرد. به پیشنهاد تورینگ، این آزمون که می‌توان به آسانی آن را اجرا کرد، مشخص می‌کند که آیا یک ماشین به حد کافی هوشمند است یا خیر.

تورینگ مقاله مورد نظر را این گونه آغاز می‌کند: «من پیشنهاد می‌کنم که این پرسش را مد نظر قرار دهید: آیا ماشین‌ها می‌توانند فکر کنند؟» سپس از آنجا که تعریف دقیق تفکر بسیار مشکل است، تورینگ پیشنهاد می‌کند که این پرسش به گونه دیگری مطرح شود: «آیا قابل تصور است که کامپیوترهای دیجیتالی بتوانند در بازی تقلید، عملکرد مناسبی از خود ارائه دهند؟» پرسشی که به گمان تورینگ دلیلی برای منفی بودن پاسخ آن وجود نداشت. در مورد شرایط دقیق آزمون تورینگ بحث‌های زیادی مطرح است که باعث شده نسخه‌های مختلفی از این آزمون به وجود آید. نکته اول شیوه انجام این آزمایش است که تقریباً همه اعتقاد دارند که نمی‌توان تنها به یک آزمایش اتکا کرد و باید درصد موفقیت در تعداد زیادی آزمایش محاسبه شود. نکته بعدی در میزان اطلاعات پیش از آزمایش داور است. به عنوان مثال، برخی پیشنهاد کرده‌اند که لزومی ندارد داور بداند یکی از افراد درگیر در آزمایش کامپیوتر است و برخی دیگر اعتقاد دارند که مشکلی با دانستن این موضوع وجود ندارد چرا که در واقع آزمون تورینگ برای توانایی فریب دادن داور طراحی نشده بلکه صرفاً سنجش میزان توانایی ماشین در شبیه‌سازی رفتارهای انسانی مدنظر است.

در اینجا باید به نکته مهمی در رابطه با آزمون تورینگ اشاره کرد. تا قبل از ارائه آزمون تورینگ، دانشمندان فعال در زمینه علوم شناختی و هوش مصنوعی مشکلات فراوانی را برای تعریف دقیق هوشمندی و مشخص کردن این که چه زمانی می‌توان یک فرآیند را تفکر نامید، تجربه می‌کردند. تورینگ که یک ریاضیدان خبره بود با ارائه آزمون تورینگ در واقع سعی داشت تا از دنیای تعاریف



نادقیقی که هضم آن برای حوزه‌های دقیقی مانند علوم کامپیوتر مشکل بود، فاصله گرفته و معیاری مشخص برای میزان هوشمندی ماشین‌ها ارائه کند. دانیل کلمنت دنت، دانشمند علوم شناختی و فیلسوف امریکایی در این رابطه می‌گوید: «هنگامی که تورینگ، آزمون مورد نظر را برای هوشمندی ماشین‌ها ارائه کرد، هدف وی بنا کردن پلتفرمی برای انجام تحقیقات علمی نبود بلکه وی آزمون تورینگ را به عنوان یک ختم‌الکلام برای بحث‌های مورد نظر در آن زمان ارائه کرد. در واقع، کلام اصلی تورینگ در مقابل کسانی که اصولاً تعریف هوشمندی برای ماشین را غیرقابل قبول می‌دانستند، این بود که: هر ماشینی که بتواند این آزمون را به صورت عادلانه‌ای پشت سر بگذارد، قطعاً یک موجود هوشمند است و دیگر بحثی در این زمینه باقی نمی‌ماند.» دنت سپس به بحث در مورد هوشمندی در قرن ۱۷ توسط دکارت اشاره می‌کند و متذکر می‌شود که وی نیز روشی مشابه برای تعریف هوشمندی ارائه داده بود که براساس برقرار کردن یک مکالمه با موجود مورد نظر بنا شده بود. در نتیجه تورینگ ادعا نمی‌کند ماشینی که نتواند با ما به شکل درستی مکالمه برقرار کند هوشمند نیست، بلکه صرفاً ادعا دارد اگر ماشینی این توانایی را داشته باشد شکی در هوشمندی آن باقی نمی‌ماند.

نام اصلی آزمون تورینگ «بازی تقلید» یا Imitation Game است. در نسخه‌ی اولیه‌ی این بازی خبری از هوش مصنوعی نبود. در این نسخه، یک داور، یک شرکت کننده‌ی مرد و یک شرکت کننده‌ی زن در سه اتاق جداگانه قرار می‌گرفته‌اند. وظیفه‌ی داور صحبت با دو شرکت کننده به صورت متنی و از طریق یک کنسول رایانه‌ای بود؛ پس از گفتگوی متنی با هر دو شرکت کننده، داور بایستی تصمیم می‌گرفت که کدام یک از شرکت کنندگان مرد است. در این بازی، هدف شرکت کننده‌ی مرد این بود که بتواند مذکر بودن خود را ثابت کند؛ هدف شرکت کننده‌ی زن نیز این بود که داور را فریب دهد و وی را متقاعد کند که او یک مرد است. اگر شرکت کننده‌ی زن موفق می‌شد داور را متقاعد کند که او در حال صحبت کردن با یک مرد است؛ وی در این بازی برنده می‌شد.

شاید بپرسید این بازی نسبتاً ساده چه ارتباطی با هوش مصنوعی دارد؟ بر اساس پیشنهاد تورینگ، می‌توان به جای قرار دادن یک زن و یک مرد در دو سوی این رقابت، یک انسان و یک رایانه را در دو سوی این رقابت قرار داد؛ در این حالت، وظیفه‌ی داور نیز شناسایی رایانه خواهد بود. به عبارت دیگر، داور به مدت پنج دقیقه به گفتگوی متنی با دو شرکت کننده (یکی انسان و دیگری رایانه) می‌پردازد و در این بین وظیفه‌ی رایانه فریب دادن داور است. برای دستیابی به نتیجه‌ی نهایی، این آزمون بارها تکرار می‌شود؛ اگر در بیش از نیمی از موارد، داور فریب خورده و رایانه را به عنوان انسان قلمداد کند، این رایانه در آزمون تورینگ موفق شده است و می‌توان آن را «هوشمند» قلمداد کرد.

### ۳-۲- آیا این هوشمند سازی است؟

در بیش از شصت سالی که آزمون تورینگ در حوزه هوش مصنوعی حضور داشته است، انتقادات مختلفی به آن وارد شده که بخش بزرگی از آن‌ها بر این موضوع استوار بوده‌اند که آیا این آزمون معیار خوبی برای تشخیص هوشمندی یک سیستم است؟

به عنوان مثال، جان سیرل فیلسوف امریکایی در مقاله‌ای با عنوان «ذهن‌ها، مغزها و برنامه‌ها» در سال ۱۹۸۰ آزمایشی ذهنی با عنوان «اتاق چینی» را طراحی کرد که به تعریف هوشمندی مورد نظر حوزه هوش مصنوعی حمله می‌کند.

فرض کنید که شما یک برنامه در اختیار دارید که می‌تواند طوری رفتار کند که زبان چینی را می‌فهمد. این برنامه یک ورودی از کاراکترهای چینی را گرفته و براساس آن‌ها خروجی متشکل از کاراکترهای چینی تولید می‌کند. همین طور فرض کنید که این برنامه آزمون تورینگ را با موفقیت پشت سر بگذارد. حال در اینجا یک پرسش بزرگ به وجود می‌آید: «آیا این ماشین به‌راستی چینی می‌فهمد یا تنها می‌تواند فهم زبان چینی را شبیه‌سازی کند؟» سیرل بیان می‌کند که اگر وی در اتاقی، مقابل این ماشین قرار بگیرد، می‌تواند با وارد کردن هر ورودی چینی در کامپیوتر و یادداشت کردن خروجی برنامه روی یک تکه کاغذ آزمون تورینگ را با موفقیت پشت سر بگذارد. وی سپس اشاره می‌کند که فرقی میان نقش ماشین در حالت اول و نقش وی در حالت دوم وجود ندارد و از آنجایی که وی یک کلمه چینی نمی‌فهمد، در نتیجه ماشین نیز درکی از زبان چینی ندارد. در نهایت وی نتیجه می‌گیرد که بدون درک شیوه عملکرد کامپیوتر و تنها از روی مشاهده رفتار آن نمی‌توان نتیجه گرفت که کاری که ماشین انجام می‌دهد فکر کردن است.

دیدگاه جان سیرل از طرف دانشمندان علوم شناختی مورد انتقادات فراوانی قرار گرفته است. از جمله این انتقادات می‌توان به این نکته اشاره کرد که ممکن است فرد به صورت خاص زبان چینی را نفهمد اما سیستم به صورت یک کل توانایی فهم زبان چینی را دارد و نمی‌توان توانایی فهم انسان به عنوان بخشی از این سیستم را از کل جدا کرد. هر چند آزمایش «اتاق چینی» مورد انتقادات فراوانی قرار گرفته و نمی‌تواند به عنوان یک خطر جدی برای آزمون تورینگ تلقی شود، اما با مشاهده چنین دیدگاه‌هایی کاملاً مشخص می‌شود که چرا پیاده‌سازی ایده آزمون تورینگ در دنیای واقعی تا این اندازه مشکل است.

دسته دیگری از انتقادات به این موضوع اشاره دارند که میزان تقلید از رفتارهای انسانی لزوماً معیار خوبی برای هوشمندی نیست. چراکه نه تمام رفتارهای انسانی هوشمندانه است و نه تمام رفتارهای هوشمندانه انسانی است. این که تا چه حد این جمله را قبول دارید، می‌تواند موضوع خوبی برای یک بحث فلسفی طولانی باشد و البته بعید است به نتیجه مشخصی برسد. به عنوان مثال، ابرکامپیوتر دیپ‌بلو ساخت آی‌بی‌ام را در نظر بگیرید که در دهه ۱۹۹۰ موفق شد گری کاسپاروف استاد مسلم شطرنج جهان را شکست دهد. دیپ‌بلو طبیعتاً نمی‌تواند در مکالمه با انسان همراهی کند اما به خوبی

وی (حتی بهتر از او) شطرنج بازی می‌کند. آیا این ماشین کمتر از الیزا هوشمند است؟ جواب از نظر بسیاری خیر است. اما باز هم باید توجه داشت که تورینگ به هیچ عنوان ادعا نمی‌کند عدم تقلید از انسان به معنای عدم هوشمندی است.

این که آیا تقلید از رفتار انسان واقعاً نشان‌دهنده هوشمندی است یا خیر، هنوز مورد بحث و بررسی است. به عبارتی، هنوز هم تعریف دقیقی برای هوشمندی در اختیار نداریم و همین موضوع باعث می‌شود تا نتوان در این مورد استدلال چندان قابل قبولی ارائه داد. به هر روی، ما امروز می‌دانیم که رفتار هوشمندانه و رفتار انسانی ممکن است لزوماً به یک معنی نباشند. همچنین آگاه هستیم که برای گذراندن آزمون تورینگ، آشنایی ماشین به جزئیات و قوانین زبان انسانی به همان اندازه اهمیت دارد که دانش و استدلال گنجانده شده در آن ارزشمند است. خبر نه‌چندان امیدوار کننده، این است که با وجود پیشرفت‌های فراوان حوزه یادگیری زبان و زبان‌شناسی، فرآیند دقیقی که باعث می‌شود انسان‌ها در یادگیری یک زبان به چنین درجه‌ای از تبحر دست‌یابند، به طور دقیق برای دانشمندان مشخص نیست. حتی از تمام این موارد که بگذریم، مسئله‌ای بسیار مهم‌تر مطرح می‌شود و آن این است که آیا اصولاً گذراندن یا نگذراندن آزمون تورینگ تا این حد مسئله مهمی است؟ دنیای نوین هوش مصنوعی اعتقاد دارد که پاسخ این پرسش منفی است. در ادامه مقاله می‌کوشیم تا تصویری از وضعیت آزمون تورینگ در دنیای امروز ترسیم کنیم.

#### ۴-۲- مراحل هوشمند سازی

در این قسمت مراحل که هوش مصنوعی می تواند به آن دست یابد را بررسی می کنیم :

انواع مراحل در هوش مصنوعی :

۱- هوش مصنوعی باریک یا ضعیف (ANI: Artificial Narrow Intelligence)

۲- هوش مصنوعی عمومی یا قوی (AGI: Artificial General Intelligence)

۳- هوش مصنوعی فوق العاده (ASI: Artificial Super Intelligence)

##### ۴-۲-۱- هوش باریک مصنوعی (ANI: Artificial Narrow Intelligence):

هوش مصنوعی باریک یا هوش مصنوعی ضعیف طبقه ای از هوش مصنوعی می باشد که ماشین می تواند فقط وظایف باریک از پیش تعیین شده را انجام دهد.

در این طبقه در حقیقت ماشین قابلیت فکر کردن را دارا نمی باشد و فقط وظایفی که از قبل برای آن تعریف شده است را انجام می دهد. نمونه ی این طبقه سیری در گوشی های ایفون می باشد. سیری یک مثال خوب از هوش مصنوعی ضعیف است چون در یک محدوده از پیش تعیین شده عمل می کند و هوشمندی حقیقی و خودآگاهی ندارد گرچه یک سیستم پیچیده دارای هوش مصنوعی ضعیف است. در سیری در عمل در یک محدود از پیش تعریف شده و محدوده وجود دارد.

##### ۴-۲-۲- هوش مصنوعی عمومی یا قوی (AGI: Artificial General Intelligence)

هوش مصنوعی عمومی که به هوش مصنوعی قوی هم معروف می باشد انقلابی کامل در هوش مصنوعی خواهد بود در این نوع دسته از هوش مصنوعی ماشین قابلیت فکر کردن و دارای قدرت تصمیم گیری می باشند دقیقاً مثل انسان. در حال حاضر انسان هنوز به این مرحله از هوش مصنوعی دست نیافته است ولی باور ها بر این است که در آینده ی نزدیک ماشینی هوشمند به اندازه انسان ساخته خواهد شد. عده ای دانشمند ها این نوع از هوش مصنوعی را تهدیدی بزرگ برای انسان ها می دانند از جمله استیون هاوکنینگ.

استیون ها فکینگ در این مورد گفته است :

" دست یابی و تکامل کامل هوش مصنوعی (دست یافن به هوش مصنوعی قوی ) میتواند صدای پایان نژاد انسان ها باشد... ین هوش می تواند با اصلاح و پیشرفت مداوم خود، نسبت به هوش انسانی برتری یابد. انسان که تکامل بیولوژیکی دارد هرگز نمی تواند حریف هوش مصنوعی بشود "

با کمی عمل می توان سوالی که در ذهن شکل میگیرد که آیا هوش مصنوعی تهدید است یا فرصت؟

### ۳-۴-۲- هوش مصنوعی فوق العاده (ASI: Artificial Super Intelligence)

این مرحله از هوش مصنوعی زمانی است هوش مصنوعی از هوش انسان پیشی میگیرد و قابلیت در دست گرفتن دنیا را دارد .

شاید در نگاه اول این نوع مرحله کمی عجیب و دور از انتظار به نظر برسد ولی الن ماسک در این مورد گفته است که سرعت پیشرفت تکنولوژی هوش مصنوعی به اندازه ی یک افتجار زیاد است و وقوع اتفاقی خطرناک در حیطه ی این موضوع در پنج سال آینده محتمل و در ده سال آینده حتمی می باشد.

بنابراین ، اینها مراحل مختلف هوشمندی بودند که یک ماشین می تواند به دست آورد. حال در این مرحله هوش مصنوعی را بر اساس عملکرد آن ها دسته بندی و مورد بحث قرار می دهیم:

### ۵-۲- انواع هوش مصنوعی :

اگر سوال شود که انواع هوش مصنوعی چیست باید در نظر گرفت که هوش مصنوعی را باید بر اساس عملکرد آن ها دسته بندی کرد .

بر اساس عملکرد سیستم های مختلف هوش مصنوعی، هوش مصنوعی می تواند به گروه های زیر دسته بندی شود:

۱- ماشین های هوش مصنوعی واکنش گر (Reactive Machines AI)

۲- هوش مصنوعی حافظه محدود (Limited memory AI)

۳- نظریه ذهن هوش مصنوعی (Theory Of Mind AI)

۴- هوش مصنوعی خودآگاه (Self-aware AI)

هر کدام از این دسته بندی ها به صورت مختصر در ادامه توضیح داده خواهد شد :

#### ۱-۵-۲- ماشین های هوش مصنوعی واکنش گر (Reactive Machines AI)

ساده ترین نوع هوش مصنوعی است و برداشتی ساده و مستقیم از داده های حال (همان لحظه) دارد و بر اساس آن چه در لحظه دریافت کرده است عمل می کند. این نوع هوش مصنوعی فاقد قابلیت پیش بینی نسبت به آینده از داده های دریافتی در لحظه را دارد. در حقیقت این نوع ماشین قابلیت اجرا کردن یک سری از دستور العمل های از پیش تعیین شده را دارد .  
نمونه ی ماشین واکنش گر برنامه ی شطرنج شرکت IBM می باشد که توانست بهترین شطرنج باز دنیا را شکست دهد.



شکل ۱ هوش مصنوعی واکنش گر

## ۲-۵-۲- هوش مصنوعی حافظه محدود (Limited memory AI)

این نوع هوش مصنوعی یک پله بالاتر از نوع واکنشگراست و با دارا بودن حافظه، داده‌های گذشته را در تصمیم‌گیری فعلی خود دخیل می‌کند. حافظه و تجربه این نوع هوش مصنوعی است به اندازه‌ای است که بتواند تصمیمات صحیح‌تری بگیرد و اعمال مناسبی را اتخاذ کند.

در حقیقت ماشین با رجوع به حافظه‌ی گذشته خود و در خیل کردن آن اطلاعات تصمیمات جدید می‌گیرد که باعث بهبود عملکرد ماشین می‌شود.

ماشین‌های خودران نمونه‌ای از این دسته می‌باشند. برای مثال ماشین هوشمند خودران با استفاده از سنسورها و اطلاعات دریافتی آن موانع، خطوط خیابان و تابلوهای نصب شده در مسیر را تشخیص می‌دهد و با تکیه بر این اطلاعات تصمیم دقیق‌تر و کامل‌تری می‌گیرد و از تصادف آینده جلوگیری می‌کند.

## ۲-۵-۳- نظریه ذهن هوش مصنوعی (Theory Of Mind AI)

هوش مصنوعی در این نوع می‌تواند احساسات و افکاری که بر رفتار انسان اثر می‌گذارند را درک کند. این هوش مصنوعی می‌تواند احساسات، انگیزه‌ها و انتظارات را درک کرده و از نظر اجتماعی فعال باشد. این نوع از هوش مصنوعی هنوز کامل شکل نگرفته است ولی پیشرفت قابل توجهی داشته است.

## ۲-۵-۴- هوش مصنوعی خودآگاه (Self-aware AI)

هوش مصنوعی خودآگاه خود دارای تصویر و تجسم است. این نوع هوش در واقعاً تکامل یافته "نظریه ذهن" است و غیر از داشتن حالات روحی نسبت به خود نیز آگاهی و اطلاع دارد. در این سطح از هوش مصنوعی ماشین می‌تواند رفتار، احساسات و عکس‌العمل دیگران را پیش‌بینی کند. شاید آخرین مرحله هوش مصنوعی همین مرحله باشد که می‌تواند جهان را دگرگون کند.

## ۶-۲- شاخه های هوش مصنوعی :

هوش مصنوعی می تواند برای حل مشکلات در دنیای حقیقی مورد استفاده قرار گیرد روش های متعددی برای اعمال هوش مصنوعی در حل مشکلات وجود دارد که در ادامه بیان می شود در حقیقت موارد زیر شاخه های هوش مصنوعی می باشند:

۱. یادگیری ماشین (Machine Learning)

۲. یادگیری عمیق (Deep Learning)

۳. پردازش زبان طبیعی (Natural Language Processing)

۴. رباتیک (Robotics)

۵. سیستم های متخصص (Expert Systems)

۶. منطق فازی (Fuzzy Logic)

هر کدام به مختصر توضیح داده خواهد شد.

### ۱-۶-۲- یادگیری ماشین (Machine Learning)

آیا تا به حال برای خرید آنلاین از وبسایت های اینترنتی اقدام کرده اید؟ اگر پاسخ شما مثبت است، حتما متوجه شده اید که هنگام جست و جوی کالای مورد نظر خود، سیستم به صورت خودکار کالاهای مشابهی را نیز توصیه می کند. همچنین ممکن است ملاحظه کرده باشید که برخی سیستم ها، به صورت خودکار خریدهای دیگر کاربرانی که کالای مورد نظر شما را خریداری کرده اند را نیز نشان می دهند و سعی در جلب توجه شما دارد. این سیستم ها، به عنوان یک ماشین، چطور این کارها را انجام می دهند؟ این، همان یادگیری ماشین است.

آرتور ساموئل (Arthur Samuel) آمریکایی، یکی از پیشروهای حوزه بازی های کامپیوتری و هوش مصنوعی، عبارت “یادگیری ماشین” را در سال ۱۹۵۹ که در IBM کار می کرد، به ثبت رساند. یادگیری ماشین، که از الگوشناسی و نظریه یادگیری محاسباتی الهام گرفته شده است، مطالعه و ساخت الگوریتم هایی را که می توانند بر اساس داده ها یادگیری و پیش بینی انجام دهند بررسی می کند – چنین الگوریتم هایی از دستورات برنامه پیروی صرف نمی کنند و از طریق مدلسازی از داده های ورودی نمونه، پیش بینی یا تصمیم گیری می کنند .



یادگیری ماشین، زیر مجموعه‌ای از هوش مصنوعی است. با استفاده از تکنیک‌های یادگیری ماشین، کامپیوتر، الگوهای موجود در داده‌ها (اطلاعات پردازش شده) را یاد گرفته و می‌تواند از آن استفاده کند. توجه داشته باشید که در این تکنیک‌ها، یادگیری در یک سیستم کامپیوتری، بدون برنامه‌نویسی صورت می‌پذیرد.

مثال کلاسیک زیر را در نظر بگیرید

فرض کنید در یک فروشگاه بزرگ خرده‌فروشی به صورت اینترنتی در حال خرید هستید. در زمان خرید، سه محصول مختلف را به سبد خرید خود اضافه می‌کنید. فرض کنید این سه محصول به صورت زیر است:

N لپ تاپ سری

موس بیسیم

یک عدد تمیز کننده مانیتور

حال، سیستم می‌خواهد به صورت هوشمند، به شما چند محصول دیگر را پیشنهاد دهد. مدل (مثلاً یک سری) برنامه‌نویسی صریح، به این صورت است که مثلاً، سیستم، محصولات هم‌دسته را به شما نمایش بدهد. در این حالت، هوشمندی خاصی (است IT محصولات) که مربوط به حوزه‌ی در سیستم مشاهده نمی‌شود و در واقع، سیستم (ماشین) یادگیری خاصی انجام نمی‌دهد.

حال فرض کنید، سیستم از طریق الگوریتم‌های یادگیری ماشین، بتواند مشتریان قبلی خود را به گفته می‌Clustering گروه‌های مختلف تقسیم‌بندی کند (به این کار به اصطلاح خوشه بندی یا شود). با این کار، شما با تکمیل سبد خرید خود، به دسته‌ای از مشتریان قبلی متعلق می‌شوید. با تعلق شما به گروه خاصی از مشتریان، محصولاتی که آن‌ها (قبلاً) خریداری کرده‌اند (و شما در سبد خرید خود ندارید) به شما پیشنهاد داده می‌شود.

مساله اصلی در یادگیری ماشین، عرضه و کلی‌سازی است. عرضه نمونه‌های داده‌ای و توابعی که بر اساس این نمونه‌ها ارزیابی می‌شوند، همگی بخشی از سیستم‌های یادگیری ماشین هستند. کلی‌سازی به معنی این قابلیت است که سیستم روی نمونه‌های داده‌ای نادیده نیز به خوبی عمل خواهد کرد. شرایطی که تحت آنها بتوان این مساله را تضمین کرد، از موضوعات اصلی مطالعه در زیرمجموعه نظریه یادگیری محاسباتی است.

## ۲-۶-۲- نحوه عملکرد یادگیری ماشین

الگوریتم‌های یادگیری ماشین، با استفاده از مجموعه داده‌هایی با عنوان داده‌های آموزشی ( training data set) یادگیری کرده و مدل‌های موردنیاز را ایجاد می‌کنند. زمانی که داده‌های جدیدی به الگوریتم یادگیری ماشین معرفی می‌شوند، سیستم می‌تواند بر اساس مدل ایجاد شده، فرایند پیش‌بینی را انجام دهد.

پیش‌بینی صورت گرفته به دقت ارزیابی شده و در صورت تایید دقت‌پذیری، الگوریتم یادگیری ماشین مذکور استقرار می‌یابد. در صورت عدم تایید دقت‌پذیری پیش‌بینی نیز، الگوریتم یادگیری ماشین با استفاده از داده‌های آموزشی کامل‌تری بارها و بارها آموزش داده می‌شود تا بتواند نتیجه مطلوب را ارائه دهد.

این، صرفاً یک مثال ایده‌آل است و در عمل، فاکتورها و مراحل بسیاری در فرایند یادگیری ماشین دخیل هستند.

یادگیری ماشین خود به سه دسته تقسیم بندی می‌شود :

- یادگیری با نظارت ( Supervised Learning )
- یادگیری بدون نظارت ( Unsupervised Learning )
- یادگیری تقویتی ( Reinforcement Learning )

در ادامه هر کدام به صورت مختصر توضیح داده خواهد شد:

## ۲-۶-۳- یادگیری با نظارت ( Supervised Learning )

یادگیری تحت نظارت را می‌توان به آموزش دانش‌آموزان تحت نظر و هدایت یک معلم تشبیه کرد. در این‌جا، مجموعه‌ای از داده‌ها را داریم که درست مثل یک معلم عمل می‌کنند و وظیفه تعلیم ماشین یا مدل را بر عهده دارند. زمانی که مدل مربوطه یادگیری کرد، قادر خواهد بود تا پیش‌بینی‌ها و تصمیمات دقیق لازم در مورد داده‌های جدید ورودی به سیستم را ارائه دهد.

اغلب روش‌های یادگیری ماشین از یادگیری نظارت شده استفاده می‌کنند. در یادگیری ماشین نظارت شده، سیستم تلاش می‌کند تا از نمونه‌های پیشینی بیاموزد که در اختیار آن قرار گرفته. به عبارت دیگر، در این نوع یادگیری، سیستم تلاش می‌کند تا الگوها را بر اساس مثال‌های داده شده به آن فرا بگیرد.

همانطور که پیش از این بیان شد، در یادگیری ماشین مجموعه داده (هایی) به الگوریتم داده می‌شود و ماشین منطق خود را بر اساس آن مجموعه داده (ها) شکل می‌دهد. این مجموعه داده دارای سطرها و ستون‌هایی است. سطرها (که از آن‌ها با عنوان رکورد و نمونه داده نیز یاد می‌شود) نماینده نمونه داده‌ها هستند. برای مثال اگر مجموعه داده مربوط به بازی‌های فوتبال (وضعیت جوی) باشد، یک سطر حاوی اطلاعات یک بازی خاص است. ستون‌ها (که از آن‌ها با عنوان خصیصه، ویژگی، مشخصه نیز یاد می‌شود) در واقع ویژگی‌هایی هستند که هر نمونه داده را توصیف می‌کنند.

در مثالی که پیش‌تر بیان شد، مواردی مانند وضعیت هوا شامل ابری بودن یا نبودن، آفتابی بودن یا نبودن، وجود یا عدم وجود مه، بارش یا عدم بارش باران و تاریخ بازی از جمله ویژگی‌هایی هستند که وضعیت یک مسابقه فوتبال را توصیف می‌کنند. حال اگر در این مجموعه داده به عنوان مثال، ستونی وجود داشته باشد که مشخص کند برای هر نمونه داده در شرایط جوی موجود برای آن نمونه خاص بازی فوتبال انجام شده یا نشده (برچسب‌ها) اصطلاحاً می‌گوییم مجموعه داده برچسب‌دار است. اگر آموزش الگوریتم از چنین مجموعه داده‌ای استفاده شود و به آن آموخته شود که بر اساس نمونه داده‌هایی که وضعیت آن‌ها مشخص است (بازی فوتبال انجام شده یا نشده)، درباره نمونه داده‌هایی که وضعیت آن‌ها نامشخص است تصمیم‌گیری کند، اصطلاحاً گفته می‌شود یادگیری ماشین نظارت شده است.

به صورت خلاصه و کلی می‌توان گفت یادگیری با نظارت برای داده‌های برچسب شده می‌باشد.

مثالی که در این مقاله مورد بحث قرار گرفته است که در فصل‌های بعدی مفصل بحث شده است استفاده از داده‌های دیجیتالی می‌باشد. داده‌ها شامل ۱۵ هزار کامنت می‌باشد کامنت‌ها به دو

دسته تقسیم شده است کامنت هایی که نشانه ی رضایت مشتری و پیشنهاد مشتری برای خرید کالا و کامنت هایی که نشانه ی عدم رضایت مشتری و عدم پیشنهاد مشتری برای خرید کالا می باشد .

در حقیقت کامنت ها برچسب شده اند حال اگر این نوع داده را در یادگیری ماشین استفاده کنیم در حقیقت یادگیری با نظارت انجام گرفته است و ماشین را قادر می سازیم با ورود داده جدید بر اساس داده های گذشته خروجی را پیش بینی نماید .

#### ۴-۶-۲- یادگیری بدون نظارت ( Unsupervised Learning )

در این حالت، مدل از طریق مشاهدات یادگیری کرده و دستورالعمل ها و ساختارهای موجود در مجموعه ی داده ها را کشف می کند. زمانی که مجموعه داده ای به مدل معرفی می شود. مدل با استفاده از خوشه بندی داده ها، ارتباطات و الگوهای موجود در آن ها را به صورت اتوماتیک کشف می کند. تنها کاری که چنین سیستمی نمی تواند انجام دهد، برچسب زنی روی دسته های مختلف است. برای مثال، با وجود این که یک سیستم یادگیری ماشین بدون نظارت قادر است دو نوع میوه سیب و انبه را به راحتی از یکدیگر سوا کند، اما نمی تواند نام آن ها را به صورت جداگانه روی هر دسته مشخص کند. درواقع به حالت ساده تر می توان گفت که در ابتدا تمامی نمونه هایی که به آن داده می شوند، هیچ برچسبی ندارند در صورتی که در یادگیری نظارتی تمامی داده ها برچسب داشتند.

یان لیوکن (Yann LeCun) ، دانشمند فرانسوی کامپیوتر و پدر بنیان گذار شبکه عصبی پیچشی (Convolutional Neural Networks | CNN)، یادگیری ماشین نظارت نشده را چنین تعریف کرده است: «آموزش دادن ماشین ها برای یادگیری برای خودشان بدون آنکه به آن ها صراحتاً گفته شود کاری که انجام می دهند درست محسوب می شود یا غلط. یادگیری نظارت نشده راهی به سوی هوش مصنوعی «حقیقی» است. »

روش های متعددی برای استفاده از حالت نظارت نشده وجود دارد که کاربردی ترین آن خوشه بندی می باشد.

خودرمز گذارها (Autoencoders)

شبکه باور عمیق (Deep Belief Network)

یادگیری هبیان/هبین (Hebbian Learning)

شبکه‌های تولید کننده رقابتی (Generative Adversarial Networks | GAN)

نقشه‌های خودسازمان دهنده (Self-Organizing maps | SOM)

باتوجه به این که موضوع مورد بحث ما خارج از این موضوع می باشد از توضیح روش های ذکر شده پرهیز می کنیم .

### ۵-۶-۲- یادگیری تقویتی (Reinforcement Learning)

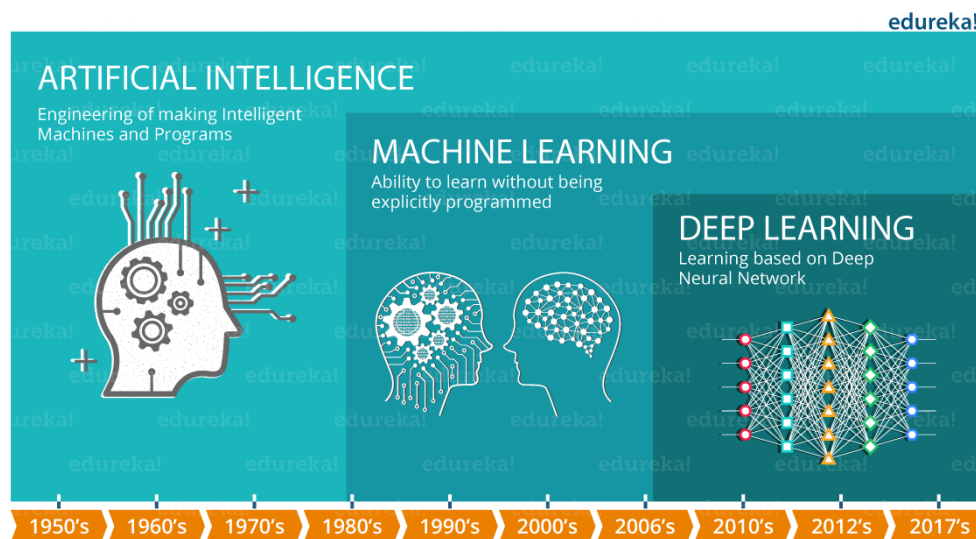
نوع سوم از الگوریتم‌ها که شاید بتوان آن‌ها را در زمره الگوریتم‌های بدون ناظر هم دسته بندی کرد. در این نوع یک ماشین (در حقیقت برنامه کنترل کننده آن)، برای گرفتن یک تصمیم خاص آموزش داده می‌شود و ماشین بر اساس موقعیت فعلی (مجموعه متغیرهای موجود) واکنش‌های مجاز (مثلاً حرکت به جلو، حرکت به عقب و ...) یک تصمیم را می‌گیرد که در دفعات اول، این تصمیم می‌تواند کاملاً تصادفی باشد و به ازای هر اکشن یا رفتاری که بروز می‌دهد، سیستم یک فیدبک یا بازخورد به او می‌دهد و از روی این فیدبک ماشین متوجه می‌شود که تصمیم درست را اتخاذ کرده است یا نه که در دفعات بعد در آن موقعیت، همان اکشن را تکرار کند یا اکشن و رفتار دیگری را امتحان کند. با توجه به وابسته بودن حالت و رفتار فعلی به حالات و رفتارهای قبلی، فرآیند تصمیم‌گیری مارکوف، یکی از مثال‌های این گروه از الگوریتم‌ها می‌تواند باشد. الگوریتم‌های شبکه‌های عصبی هم می‌توانند از این دسته به حساب آیند. منظور از کلمه تقویت شونده در نام گذاری این الگوریتم‌ها هم اشاره به مرحله فیدبک و بازخورد است که باعث تقویت و بهبود عملکرد برنامه و الگوریتم می‌شود. یادگیری تقویتی نیز به توانایی ارتباط یک عامل با محیط خارجی به منظور دست‌یابی به بهترین نتیجه اطلاق می‌شود. مفهومی که از آن، با عنوان مدل سعی و خطا نیز یاد می‌شود. این عامل، بر اساس نتایج صحیح یا اشتباهی که به دست می‌آورد، امتیاز مثبت کسب کرده یا جریمه می‌شود و در نهایت، مدل قابلیت بهبود از طریق امتیازات مثبت و نتایج مطلوب کسب‌شده را به دست می‌آورد. این یادگیری و بهبود ادامه پیدا می‌کند تا زمانی که سیستم بتواند پیش‌بینی‌ها و تصمیمات دقیق مورد نیاز در مورد داده‌های جدید ورودی را ارائه دهد.

می‌توان گفت یادگیری تقویتی مانند یادگیری مبتنی بر آزمون و خطای انسان است. یادگیری تقویتی برای ایجاد استراتژی‌ها کاربرد دارد. کاربرد یادگیری تقویتی در آموزش بازی‌ها به رایانه‌ها است. برای نمونه می‌توان شرکت DeepMind را نام برد که در سال ۲۰۱۴ توسط گوگل خریداری شد. این شرکت تلاش می‌کند تا به الگوریتم خود، بازی قدیمی و معروف آتاری (Atari) را آموزش دهد. آلفاگو (AlphaGo) سامانه هوش مصنوعی که توسط گروه DeepMind گوگل برای انجام بازی Go طراحی شده توانست قهرمان جهانی این بازی را شکست دهد.

## ۶-۶-۲- یادگیری عمیق (Deep Learning)

یادگیری عمیق جهشی بزرگ در هوش مصنوعی بود امروزه شرکت‌های بزرگی همچون گوگل از یادگیری عمیق برای تشخیص صدا و تصویر و شرکت نتفلیکس برای بررسی رفتار مشتری استفاده می‌کند.

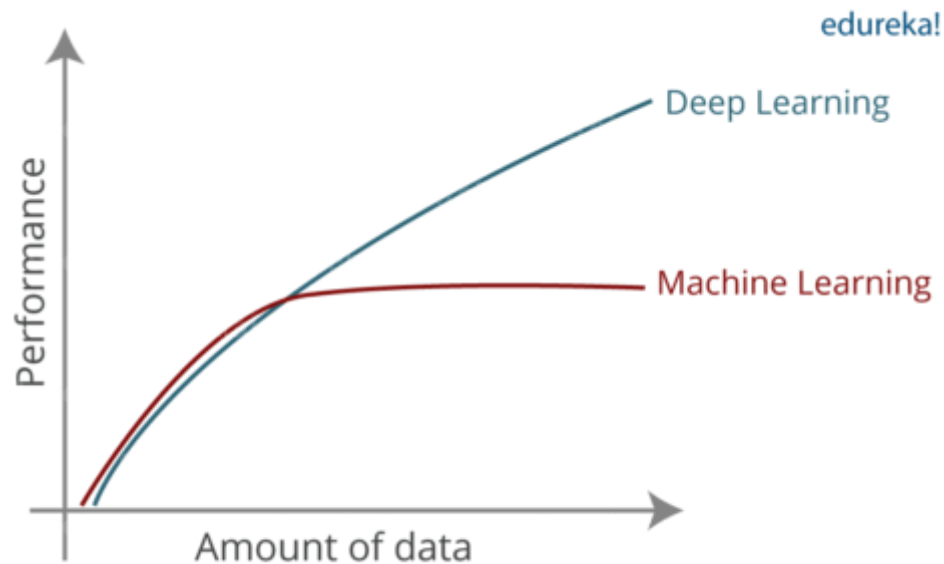
برای درک بهتر یادگیری عمیق ابتدا لازم است نگاهی به تصویر مقابل انداخت.



شکل ۲ یادگیری عمیق

باتوجه به شکل میتوان دریافت که یادگیری ماشین زیرمجموعه‌ای از هوش مصنوعی می‌باشد و یادگیری عمیق زیرمجموعه‌ای از یادگیری ماشین می‌باشد در سال‌های ظهور کرده است. در حقیقت هدف از نمایش این تصویر این بوده است که یادگیری عمیق رابطه تنگاتنگی با یادگیری ماشین دارد و در حقیقت قابلیت این را دارد که دقت یادگیری را افزایش دهد.

در شکل زیر خلاصه ی جمله قبل دیده می شود.



شکل ۳ مقایسه یادگیری ماشین و یادگیری عمیق

در یادگیری ماشین دو هدف اصلی وجود دارد

۱- کاهش میزان خطا

۲- افزایش دقت پیشبینی

اما دو مشکل اساسی که در یادگیری ماشین وجود دارد:

۱- در زمان هایی که ابعاد داده ها زیاد است (high dimensional data) در واقع زمانی است که تعداد ورودی ها و خروجی ها زیاد است یادگیری ماشین بی استفاده می شود و مدل خوبی را تحویل نمی دهد.

۲- دومین مشکل بزرگ و قابل تحمل این است که چطور می شود به کامپیوتر گفته شود که کدام قسمت داده نقش مهم تری دارد! در حقیقت اگر در هوشمند سازی کامپیوترها آن ها را قادر به پیدا کردن قسمت های مهم داده بکنیم نتیجه ای که

حاصل می شود این است که مدل ما داری دقت بیشتر و خطای کمتری خواهد بود. به این اصل استخراج ویژگی ( Feature Extraction ) گفته می شود.

در حقیقت تزریق داده ها به صورت سطری به الگوریتم نتیجه ی قابل قبولی به ما نمی دهد و استخراج ویژگی ها می تواند تحولی بزرگ در این موضوع باشد.

یادگیری عمیق (deep learning) تنها روشی است مشکل دوم، استخراج ویژگی ها (feature extraction) را حل کرده است. علت این موضوع این است که یادگیری عمیق قابلیت این را دارا می باشد که بر روی ویژگی های مهم تمرکز کند این در حالی است که با کم ترین راهنمایی در برنامه نویسی صورت میگیرد.

یادگیری عمیق شبیه سازی عملکرد مغز انسان می باشد که یادگیری آن از تجربه کردن های مغز انسان می باشد در حقیقت تجربه است که مغز انسان را آموزش می دهد. همانطور که میدانید مغز انسان از میلیون ها نورون شکل گرفته است که انسان ها را قادر به انجام کار های شکست انگیز میکند حتی مغز کودک یک ساله قادر به حل مسایل پیچیده می باشد. برای مثال:

- کودک قادر است چهره ی پدر و مادر خود را تشخیص است و تفاوت اجسام را نیز درک کند.
- کودک قادر است صداهای مختلفی از خود تولید کند و حتی قادر است افراد را از صدای آن ها به صورت متمایز تشخیص دهد.

در حقیقت مغز انسان ها به صورت ناخودآگاه خود را از ابتدای تولد آموزش می دهد. حال سوال این است که یادگیری عمیق چگونه عملکرد مغز انسان را تقلید می کند؟

یادگیری عمیق در حقیقت از نورون های مصنوعی شکل گرفته است همانطور که مغز انسان از نورون های بیولوژیکی شکل گرفته است. در حقیقت می توان گفت که یادگیری عمیق الگوریتمی را پیاده سازی می کند که از ساختار و عملکرد مغز انسان ها الهام گرفته است.

این نوع یادگیری شبکه های عصبی مصنوعی (Artificial Neural Networks) نام دارد.

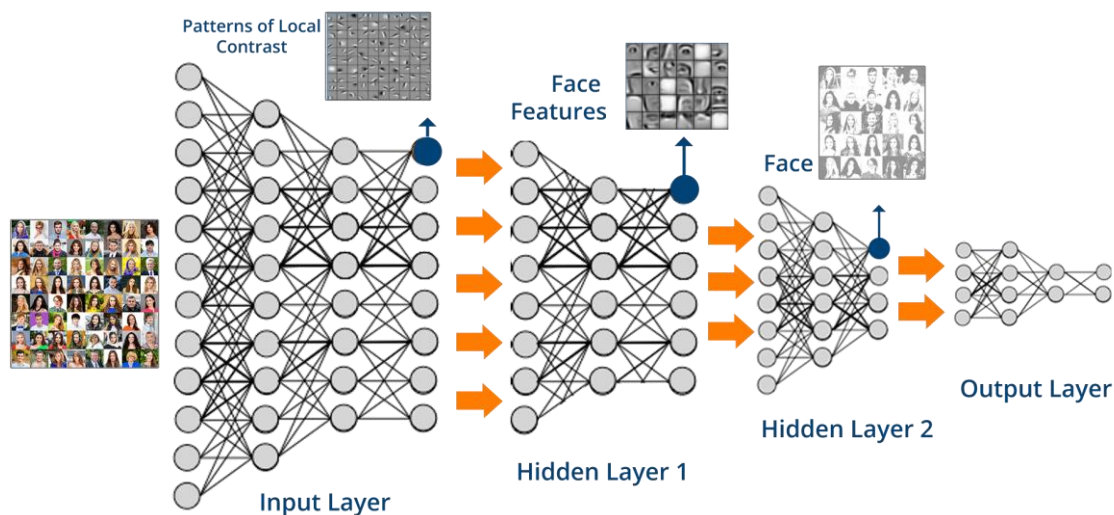
شکل کلی این الگوریتم به صورت گراف هایی متصل به هم باشد.

حال برای درک بهتر موضوع مثالی را بیان می کنیم.



فرض کنید قصد داریم سیستمی طراحی کنیم که صورت انسان های متفاوت را در عکس تشخیص دهد. اگر بخواهیم این سیستم را با یادگیری ماشین طراحی کنیم نیاز است که ویژگی هایی مانند چشم، بینی، دهان و... را تعریف کنیم تا سیستم قادر باشد ویژگی های تعریف شده را به عنوان ویژگی مهم از تصویر استخراج کند.

این در حالی است یادگیری عمیق یک مرحله از یادگیری ماشین در حالت عادی جلوتر است. در یادگیری ماشین به صورت خودکار قادر است ویژگی های مهم را برای دسته بندی کردن استخراج کند. این ویژگی بسیار مهم یادگیری عمیق نتیجه ی شبکه عصبی سیستم است.



شکل ۴ یادگیری عمیق

شکلی که در بالا مشاهده می کنید نحوه ی کار شبکه عصبی را به صورت مراحل زیر نشان می دهد:

- در ابتدایی ترین مرحله لای ورودی (سمت چپ تصویر) شبکه بر روی الگوی کنتراست مرحله ای متمرکز شده است.
- در لایه ی اول الگوهایی برای خارج کردن ویژگی های صورت قرار داده شده است.
- در لایه اخر ویژگی های صورت بر روی قالب ها اعمال می شوند.

یادگیری عمیق انواع و کاربرد های زیادی دارند دو الگوریتم مهم و کاربردی یادگیری عمیق عبارت اند از :

## ۱- RNN : Recurrent Neural Networks

## ۲- LSTM: Long-Short Term Memory

این دو الگوریتم در تولید متن (Generating Text) نقش مهم و اساسی دارند .  
در فصل های آینده به صورت مفصل در مورد این موضوع صحبت خواهد شد.

## ۷-۶-۲- پردازش زبان طبیعی ( Natural Language Processing )

پردازش زبان های طبیعی یکی از زیرشاخه های بااهمیت در حوزه گسترده علوم رایانه، هوش مصنوعی، که به تعامل بین کامپیوتر و زبان های (طبیعی) انسانی می پردازد؛ بنا بر این پردازش زبان های طبیعی بر ارتباط انسان و رایانه، متمرکز است. پس چالش اصلی و عمده در این زمینه درک زبان طبیعی و ماشینی کردن فرایند درک و برداشت مفاهیم بیان شده با یک زبان طبیعی انسانی است. به تعریف دقیق تر، پردازش زبان های طبیعی عبارت است از استفاده از رایانه برای پردازش زبان گفتاری و زبان نوشتاری. بدین معنی که رایانه ها را قادر سازیم که گفتار یا نوشتار تولید شده در قالب و ساختار یک زبان طبیعی را تحلیل و درک نموده یا آن را تولید نمایند.

در یک جمله پردازش زبان طبیعی راه ارتباطی و تعامل با کامپیوتر ها را فراهم میکند.

کاربرد های پردازش زبان بسیار گسترده است . نمونه ی این موضوع استفاده توییت از الگوریتم پردازش زبان طبیعی برای فیلتر کردن جملات و متن های خاص و یا استفاده از این الگوریتم در نظرات کاربران در سایت آمازون.

با توجه به این که موضوع اصلی این مقاله پردازش زبان طبیعی در زبان فارسی می باشد فصل بعدی مقاله را به این موضوع اختصاص داده ایم و مفصل در فصل بعدی در مورد این موضوع بحث خواهد شد.

## ۸-۶-۲- رباتیک (Robotics)

روباتیک شاخه ای از فناوری است که به طراحی، ساخت، عملیات و کاربرد روبات ها و سیستم های کامپیوتری برای کنترل، فیدبک حسگرها و پردازش اطلاعات می پردازد. این فناوری ها با دستگاه های خودکاری سر و کار دارند که می توانند جانشین انسان در محیط ها یا روندهای تولیدی خطرناک شوند یا ظاهر، رفتار و درک انسانی را شبیه سازی کنند. بسیاری از روبات های امروزی از طبیعت الهام گرفته اند که به شاخه روباتیک ملهم از بیولوژی مربوط می شوند. مفهوم ایجاد ماشین هایی که بتوانند

خودکار کار کنند، به زمان‌های دور برمی‌گردد اما تحقیق روی عملیاتی کردن و کاربردهای احتمالی روبات‌ها از قرن بیستم آغاز شد. در طول تاریخ، روبات‌ها به تقلید رفتار انسانی شناخته شده و توانسته‌اند کارهای مشابهی نیز انجام دهند. امروزه و با پیشرفت فناوری، رشته روباتیک با سرعت زیادی در حال پیشرفت است. تحقیق، طراحی و ساخت روبات‌های جدید با اهداف کاربردی متفاوت عمومی، تجاری یا نظامی انجام شده است. بسیاری از روبات‌ها کارهایی را انجام می‌دهند که برای انسان خطرناک است؛ مانند خنثی‌سازی بمب و مین و بازرسی لاشه کشتی.

#### ۹-۶-۲- سیستم‌های متخصص (Expert Systems)

در هوش مصنوعی، یک سیستم خبره (Expert System) یک سیستم کامپیوتری است که توانایی تصمیم‌سازی یک انسان خبره را شبیه‌سازی می‌کند. سیستم‌های خبره برای حل مشکلات پیچیده از طریق استنتاج در دانش خبرگی همانند یک انسان خبره است نه پیروی از دستورالعمل‌های برنامه‌نویسی. به تعریفی دیگر سیستم‌های خبره، برنامه‌های کامپیوتری‌ای هستند که نحوه تفکر یک متخصص در یک زمینه خاص را شبیه‌سازی می‌کنند. در واقع این نرم‌افزارها، الگوهای منطقی‌ای را که یک متخصص بر اساس آن‌ها تصمیم‌گیری می‌کند، شناسایی می‌نمایند و سپس بر اساس آن الگوها، مانند انسان‌ها تصمیم‌گیری می‌کنند. اولین سیستم‌های خبره در دهه ۱۹۷۰ ایجاد شدند و در دهه ۱۹۸۰ توسعه یافتند. در دهه ۱۹۷۰، ادوارد فیگن بام در دانشگاه استنفورد به دنبال کشف روش حل مسئله‌ای بود که خیلی کلی و همه منظوره نباشد. پژوهشگران دریافتند که یک متخصص معمولاً دارای شماری رموز و فوت و فن خاص برای کار خود می‌باشد و در واقع از مجموعه‌ای از شگردهای سودمند و قواعد سرانگشتی در کار خود بهره می‌برد، این یافته مقدمه پیدایش سامانه خبره بود. سامانه خبره با برگرفتن این قواعد سر انگشتی از متخصصین و به تعبیری با تبدیل فرایند استدلال و تصمیم‌گیری متخصصین به برنامه‌های رایانه‌ای می‌تواند به عنوان ابزار راهنمای تصمیم‌گیری در اختیار

غیرمتخصص و حتی متخصصین کم تجربه قرار گیرد . سیستم‌های خبره از اولین اشکال واقعا موفق نرم‌افزارهای هوش مصنوعی بودند .

هر سیستم خبره از دو بخش مجزا ساخته شده است: پایگاه دانش و موتور تصمیم‌گیری. پایگاه دانش یک سیستم خبره از هر دو نوع دانش مبتنی بر حقایق (factual) و نیز دانش غیرقطعی (heuristic) استفاده می‌کند. Factual knowledge، دانش حقیقی یا قطعی نوعی از دانش است که می‌توان آن را در حیطه‌های مختلف به اشتراک گذاشت و تعمیم داد؛ چراکه درستی آن قطعی است.

در سوی دیگر، Heuristic knowledge قرار دارد که غیرقطعی‌تر و بیشتر مبتنی بر برداشت‌های شخصی است.

سیستم‌های خبره مبحثی است مفصل که در حیطه ی این مقاله نمی باشد.

#### ۱۰-۶-۲- منطق فازی ( Fuzzy Logic)

در منطق غیرفازی تنها دو ارزش درست (true) یا نادرست (false) وجود دارد. چنین منطقی نمی‌تواند چندان کامل باشد؛ چراکه فهم و پروسه تصمیم‌گیری انسان‌ها در بسیاری از موارد، کاملا قطعی نیست و بسته به زمان و مکان آن، تا حدودی درست یا تا حدودی نادرست است. در خلال سال‌های ۱۹۲۰ و ۱۹۳۰، Jan Lukasiewicz فیلسوف لهستانی منطقی را مطرح کرد که در آن ارزش یک قانون می‌تواند بیشتر از دو مقدار ۰ و ۱ یا درست و نادرست باشد. سپس پروفیسور لطفی‌زاده نشان داد که منطق Lukasiewicz را می‌توان به صورت “درجه درستی” مطرح کرد. یعنی به جای این که بگوییم: “این منطق درست است یا نادرست؟” بگوییم: “این منطق چقدر درست یا چقدر نادرست است؟” از منطق فازی در مواردی استفاده می‌شود که با مفاهیم مبهمی چون “سنگینی”، “سرما”، “ارتفاع” و از این قبیل مواجه شویم. این پرسش را در نظر بگیرید: “وزن یک شیء ۵۰۰ کیلوگرم است، آیا این شیء سنگین است؟” چنین سوالی یک سوال مبهم محسوب می‌شود؛ چراکه این سوال مطرح می‌شود که “از چه نظر سنگین؟” اگر برای حمل توسط یک انسان بگوییم، بله سنگین است. اگر برای حمل توسط یک اتومبیل مطرح شود، کمی سنگین است، ولی اگر برای حمل توسط یک هواپیما مطرح شود سنگین نیست.

در اینجاست که با استفاده از منطق فازی می‌توان یک درجه درستی برای چنین پرسشی در نظر گرفت و بسته به شرایط گفت که این شیء کمی سنگین است. یعنی در چنین مواردی گفتن این که این شیء سنگین نیست

(false) یا سنگین است (true) پاسخ دقیقی نیست.

مثال کاربردی منطق فازی در کنترل ماشین‌های خود ران می‌باشد. در قسمت کاهش سرعت ماشین اگه از منطق غیر فازی استفاده شود دو حالت برای سیستم تعریف می‌شود

۱- زمانی که مانع رو به رو نزدیک است (true)

۲- زمانی که مانع روبه رو دور است (false)

بر فرض مثال تعریف می‌شود اگر فاصله کمتر از ۳ متر بود ترمز شود اگر بیشتر از سه متر بود ماشین در حرکت باشد در این حالت فقط از حالت true و false استفاده شده است و واضح است که نتیجه ی مطلوبی را نمی‌دهد حال اگر این سیستم با منطق فازی طراحی شود، سیستم بر اساس میزان درست یا میزلن غلط بودن ترمزها را تنظیم می‌کند نه صرفاً بر اساس درست و غلط بودن .

در این فصل تلاش شد که نگاهی کلی بر هوش مصنوعی، انواع آن و دسته بندی آن شود .

هوش مصنوعی زمینه ی بسیار گسترده است و صرفاً نمی‌توان به یک فصل از این مقاله اکتفا کرد ولی باتوجه به موضوع اصلی مقاله، پردازش زبان طبیعی در زبان فارسی ، سعی شد تعریف و اصلاحات کاربردی مورد استفاده در این زمینه بیان شود.

در ادامه نگاه دقیق تری به هوش مصنوعی در پردازش زبان می‌کنیم و موضوع را کمی تخصصی تر ادامه می‌دهیم.

## فصل سوم

### پردازش زبان طبیعی

#### ۳-۱- پردازش زبان طبیعی

دست یابی به هوش مصنوعی مستلزم آن است که روش های درک و فهم انسان ها کشف شود در حقیقت با الگو برداری از روش درک و یادگیری انسان دست یافتن به هوش مصنوعی هموار تر خواهد شد. یکی از مهم ترین و اساسی ترین موضوع در هوش مصنوعی این است که چگونه میتوان ماشینی ساخت که قادر باشد همانند انسان درک داشته باشد و بتواند ارتباط برقرار کند .

از زمان پیدایش انسان، انسان سعی کرده است که روشی برای ارتباط با انسان دیگر پیدا کند که این امر به مرور زمان باعث پیدایش زبان شده است. در حقیقت زبان وسیله ای قدرتمند در برقراری ارتباط انسان ها با یکدیگر می باشد. در نتیجه برای ایجاد یک هوش مصنوعی نیاز است که ماشین ها را مجهز به برقراری ارتباط با انسان ها بکنیم این به این معنا می باشد که باید درک زبان و استفاده از آن را جهت برقراری ارتباط برای هوشمند سازی ماشین ها به کار برد . مشکل اصلی آن (از نگاه کامپیوتر) این است که کامپیوتر به صورت پیش فرض فقط ۰ها و ۱ها را می فهمد. یعنی یک کامپیوتر نمی تواند زبان طبیعی محاوره ای ما را متوجه شود. برای همین نیاز است تا یک مجموعه عملیات (Processes) بر روی این زبان طبیعی انجام شود (Natural Language Processing) تا بتوان آن را برای کامپیوتر قابل فهم کرد.

پردازش زبان طبیعی (NLP: Natural Language Processing) شاخه ای از هوش مصنوعی می باشد که در راستای تحقق این موضوع پا به عرصه نهاده است .

در نگاه کلی میتوان گفت پردازش زبان طبیعی به دنبال برقراری ارتباط بین رایانه ها و انسان از طریق زبان طبیعی می باشد.

پس چالش اصلی و عمده در این زمینه درک زبان طبیعی و ماشینی کردن فرایند درک و برداشت مفاهیم بیان شده با یک زبان طبیعی انسانی است. به تعریف دقیق تر، پردازش زبان های طبیعی عبارت است از استفاده از رایانه برای پردازش زبان گفتاری و زبان نوشتاری. بدین معنی که رایانه ها را قادر سازیم که گفتار یا نوشتار تولید شده در قالب و ساختار یک زبان طبیعی را تحلیل و درک نموده یا آن را تولید نمایند.

پردازش زبان طبیعی محدودیت های اساسی دارد و این محدودیت ها هنوز به صورت کامل بر طرف نشده اند ولی دست یافتن به درصد عمده ای از این مسر ممکن شده است. محدودیت های اساسی پردازش زبان طبیعی :

- نیاز به درک معانی: رایانه برای آن که بتواند برداشت درستی از جمله ای داشته باشد و اطلاعات نهفته در آن جمله را درک کند، گاهی لازم است که برداشتی از معنای کلمات موجود در جمله داشته باشد و تنها آشنایی با دستور زبان کافی نباشد. مثلاً جمله محمد کتاب را نخرید زیرا گران بود. و جمله محمد کتاب را نخرید چون بی سواد بود. ساختار دستوری کاملاً یکسانی دارند و تشخیص این که کلمات «گران» و «بی سواد» به «محمد» برمی گردند یا به «کتاب»، بدون داشتن اطلاعات قبلی درباره ماهیت «حسن» و «سیب» ممکن نیست.
- دقیق نبودن دستور زبان ها: دستور هیچ زبانی آن قدر دقیق نیست که با استفاده از قواعد دستوری همیشه بتوان به نقش هریک از اجزای جمله های آن زبان پی برد.
- استفاده از اطلاعات، ضرب المثل و کنایه در زبان ها مثلاً ضرب المثل موش تو سوراخ نمیرفت جارو به دمش میبست برای رایانه ها کامل بی مفهوم است

درست است که هنوز دست یابی به پردازش زبان طبیعی به صورت کامل محقق نشده است ولی هم کیزان در دستری باعث حل بسیاری از مشکلات مهم شده است. برای مثال روزانه روزنامه ها و متن های زیادی در صفحه هات مجازی آپلود می شود پردازش زبان طبیعی دسته بندی متون را اسان کرده است و در کمترین زمان بدون نیاز به نیروی انسانی متون را موضوع بندی و دسته بندی میکند.

مزیت اساسی و مهم هوش مصنوعی این می باشد که با مرور زمان و پردازش داده های جدید و بیشتر هوش خود را ارتقا میدهد دقیقاً مانند نوزادی که به مرور زمان و حضور در محیط های مختلف مطالب

جدید یاد می‌گردد. این موضوع باعث شده است هوش مصنوعی روز به روز دقیق تر و کامل تر شود به خصوص در حوزه ی یادگیری زبان.

روش های متعددی برا پرادرش زبان طبیعی وجود دارد. برای جلوگیری از افزایش حجم مطالب و سردرگمی فقط روش های استفاده شده در این مقاله مورد بحث قرار می‌گیرد . در ابتدا لازم است که هدف مقاله را به صورت دقیق تری بررسی کنیم و متوجه آن بشویم که پردازش زبان قرار است چه مشکلی را برای ما حل کند.

هدف این مقاله این شامل سه قسمت می باشد قسمت اول مقاله قرار است به کمک هوش مصنوعی و الگوریتم های پردازش زبان طبیعی شبکه ای طراحی کنیم که کامنت ها را به صورت ورودی در نظر گرفته و خروجی آن تشخیص مثبت یا منفی بود کامنت باشد. مثبت یا منفی بودن کامنت به منظر این است که این کاربر خرید این کالا را پیشنهاد میکند یا کاربر خرید کالا را پیشنهاد نمی کند.

در قسمت دوم مقاله قرار است شبکه ای طراحی شود که نظر کاربر را تحلیل کند و خروجی که به ما می دهد این است که کاربر در مورد چه دسته کالایی نظر داده است که شامل سه دسته موادغذایی، لوازم برقی و لوازم آرایشی می باشد.

در قسمت سوم مقاله از الگوریتمی مشابه قسمت سوم استفاده می شود و خروجی آن است که اگر نظر کاربر در مورد کالای خریداری شده مثبت است ویژگی مثبت کالا از نگاه کاربر چیست و اگر نظر کاربر در مورد کالای خریداری شده منفی است ویژگی منفی کالا از نگاه کاربر چیست.

برای دست یابی به هدف اول روش های متعددی وجود دارد دو روش اساسی برای حل این موضوع مورد توجه واقع شده است که عبارت اند از :

۱- طبقه بندی متن (Text classification)

۲- تحلیل احساساتی (Sentiment Analysis)

و برای دست یابی دو هدف دیگر روشی که استفاده می شود عبارت است از :

• مدل سازی مبحث (Topic Modeling)

هرکدام از این روش ها به صورت مختصر در ادامه ی همین فصل و به صورت مفصل در فصل های بعدی توضیح داده خواهد .



دو نرم افزاری که در این زمینه مورد استفاده قرار میگیرد عبارت اند از R, Python در این مقاله از نرم افزار python استفاده شده است.

در فصل های بعدی علاوه بر توضیحات ، قسمتی کدنویسی در پایتون هم اضافه شده است. داده هایی که قرار است مورد بررسی قرار گیرد داده های آزاد شده ی سایت دیجیکالا می باشد این داده ها شامل ۱۵ هزار کامنت از کالای مختلف می باشد.

نکته قابل توجهی که وجود دارد این است دسته ای از کلمات در هر زبان وجود دارد که در تحلیل متن ها در پردازش طبیعی بی استفاده می باشد مانند کلماتی مثل :این، آن، و، یا، که، با، ایموجی ها و... این دسته از کلمات به عنوان کلمات وقفه در فارسی (Stop words in Persian) شناخته می شوند که قبل از هر پردازشی لازم که این کلمات از متن و نظرات حذف شود تا پردازش متن دقیق تر و سریع تر صورت گیرد.

در ادامه نگاه کلی و اجمالی به سه روش بیان شده و تفاوت آن ها می اندازیم و در سه فصل بعدی مفصل مورد بحث قرار میدهیم .

### ۳-۱-۱- طبقه بندی متن (Text classification)

در این روش از تحلیل با تبدیل متن به دسته ای از بردار شبکه را آموزش می دهیم .در حقیقت متن را به دسته ای از اعداد قابل فهم برای شبکه می کنیم و شبکه نسبت به دست ای از اعداد آموزش داده می شود .

### ۳-۱-۲- تحلیل احساساتی (Sentiment Analysis)

بطور کلی بررسی ارتباط میان واژه و معنا را معناشناسی می گویند. در منطق نیز بررسی ارتباط میان نمادها و آنچه که نمادها نشان می دهند را معناشناسی (Semantics) می نامند. در حقیقت در این نوع پردازش معنای جمله مورد تحلیل واقع می شود برای مثال: من خرید این اسپیکر را اصلا پیشنهاد نمی کنم چون صدای بسیار ضعیفی دارد. در تحلیل معناشناسی جمله ی بالا به علت وجود کلمات اصلا، نمی کنم ،بسیار و ضعیف جمله دارای بار معنایی منفی میباشد در مقابل نظر زیر  
من خرید این اسپیکر را حتما پیشنهاد می کنم چون صدای بسیار قوی دارد.

در تحلیل معناشناسی به علت وجود کلمات حتماً می‌کنم، بسیار و قوی جمله دارای بار معنایی مثبت می‌باشد.

پردازش زبان طبیعی به خصوص در زبان انگلیسی پیشرفت شایان و قابل توجهی داشته است

### ۳-۱-۳- مدل سازی مبحث (Topic Modeling)

ساده ترین راه برای پیدا کردن موضوع یک متن خواندن آن متن است ولی اگر تعداد داده ها زیاد باشد این راه بدون شک هوشمندانه نمی باشد برای حل این موضوع هوش مصنوعی مبحث topic modeling را بیان کرده است. الگوریتم‌های Topic Modeling به این صورت عمل می‌کنند که با مشاهده‌ی تمامی متون، سعی در ایجاد گروه‌هایی دارد که این گروه‌ها از کلمات نزدیک به هم تشکیل شده‌اند. به این گروه‌ها، موضوعات (Topics) می‌گویند.

پراکاربردترین روش که در این زمینه استفاده می‌شود روش **Latent Dirichlet Allocation** یا به اختصار LDA می‌باشد. این روش برای هر متن موضوعی و برای هر موضوعی دسته کلمه‌ای به وجود می‌آورد که به توزیع دیریکله Dirichlet Distribution شناخته می‌شود.

اگر بخواهیم به صورت کلی به موضوع نگاه کنیم مدلی ساخته می‌شود که برای هر متن باتوجه به میزان تکرار کلمات آن متن پرتکرارترین کلمات آن را به عنوان کلمات محوری آن موضوع در اختیار ما قرار میدهد. حال متن جدید پس از ورود به شبکه بررسی می‌شود که کدام دسته از کلمات در موضوع بیشترین تکرار را دارد و متن جدید به کدام موضوع نزدیک تر می‌باشد.

در این فصل سعی شد نگاه کلی به موضوع پردازش زبان و روش‌های مورد هدف مقاله بشود هر کدام از روش‌ها به صورت کامل تر در فصل‌های بعدی بررسی می‌شود و برنامه‌ی آن را در پایتون مورد بحث قرار می‌دهیم.

## فصل چهارم

### طبقه بندی متن

#### ۱-۴ - طبقه بندی متن (Text classification)

در فصل های گذشته به صورت کامل در مورد هوش مصنوعی صحبت شد در این فصل قرار است موضوعات زیر مورد بحث قرار گیرد:

- مروری کلی بر مباحث پایه ای یادگیری ماشین
- معیارهای طبقه بندی (Classification Metrics)
- ماتریس پیرشانی (Confusion Metrix)
- استخراج ویژگی های متن (Text Feature Extraction)
- آشنایی با Sckit-learn در Python و اعمال آن بر داده های واقعی

در ادامه در ابتدا با چهار مبحث اولیه شروع میکنیم و مفاهیم کلی آن را مورد بحث قرار می دهیم و در آخر با به کارگیری آن ها در پایتون بر روی داده های واقعی خروجی و نتیجه را مشاهده میکنیم. برای فهم بهتر موضوع مثالی از داده هایی که قرار اس ت در این مقاله مورد استفاده قرار گیرد استفاده می شود.

## ۲-۴- مروری کلی بر مباحث پایه ای یادگیری ماشین

همانطور که در فصل اول اشاره شد یادگیری ماشین دارای دو دسته می باشد :

۱. یادگیری با نظارت (Supervised Learning)

۲. یادگیری بدون نظارت (Unsupervised Learning)

در طبقه بندی متن از حالت اول یادگیری با نظارت استفاده می شود در این حالت داده ها دارای برچسب می باشند (Labeled) در حقیقت در این حالت داده های ورودی ما مشخص و خروجی مطلوب آن ها نیز مشخص می باشد. برای مثال:

- من خیلی از خرید این دوربین راضی ام کیفیت تصویر و رنگش عالیه خریدش پیشنهاد میکنم
  - من اصلا خرید این دوربین پیشنهاد نمی کنم خیلی سنگین و اصلا خوش دست نیست
- در بالا دو نظر از دو خریدار را مشاهده می کنید که کاربر اول خرید کالا را پیشنهاد میکند و کاربر دوم خرید کالا را پیشنهاد نمی کند.
- در داده های برچسب شده با توجه به هدف استفاده از داده و معیار ها، داده ها برچسب زده می شود در اینجا قرار است داده ها به دو دسته تقسیم شوند:

۱. نظراتی که کاربر خرید کالا را پیشنهاد می دهد (Recommended)

۲. نظراتی که کاربر خرید کالا را پیشنهاد نمی دهد (Not-recommended)

برای مثال بیان شده نظر اول برچسب پیشنهاد می شود و نظر دوم برچسب پیشنهاد نمی شود را میگیرد

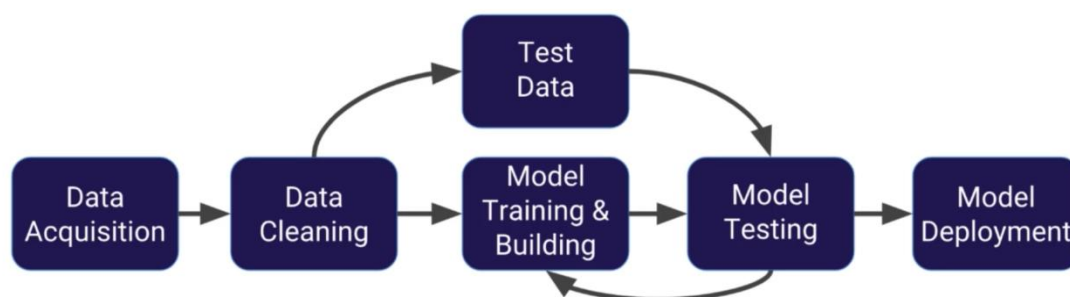
- من خیلی از خرید این دوربین راضی ام کیفیت تصویر و رنگش عالیه خریدش پیشنهاد میکنم <<<<<< پیشنهاد میشود
- من اصلا خرید این دوربین پیشنهاد نمی کنم خیلی سنگین و اصلا خوش دست نیست <<<<<< پیشنهاد نمی شود

حال داده ها ما ۱۵ هزار نظر کاربر برچسب شده می باشد که قرار است الگوریتمی به کمک یادگیری ماشین طراحی شود که مرحله ی یادگیری را به کمک این ۱۵ هزار نظر کاربر انجام دهد و شبکه ای

در اختیار ما قرار دهد که با ورودی داده ای جدید خروجی مطلوب را پیش بینی و در اختیار ما قرار دهد.

### ۴-۳-۴- مراحل یادگیری ماشین با نظارت (Unsupervised Learning)

موضوعی که در فصل های قبل بیان نشد پروسه ی یادگیری ماشین می باشد در اینجا باتوجه به کاربرد مورد استفاده مراحل یادگیری ماشین در حالت یادگیری با نظارت مورد بررسی قرار میگیرد. شکل کلی مراحل یادگیری با نظارت به صورت شکل زیر می باشد:



شکل ۵ مراحل یادگیری با نظارت

در ادامه هر کدام از مراحل بالا را به صورت جداگانه توضیح می دهیم.

#### ۴-۳-۴-۱- اکتساب داده ها (Data Acquisition)



شکل ۶ اکتساب داده ها

در مرحله اول نیاز است که داده ها تهیه شوند این داده ها غالباً در فرم های csv یا xls می باشند. داده هایی که در اینجا استفاده شده است داده های دیجیکالا می باشد که در اکسل جدول بندی شده است.

## ۲-۳-۴- تمیز کردن داده ها (Data Cleaning)



شکل ۷ تمیز کردن داده ها

داده های تهیه شده در مرحله قبل غالباً داده های نامرتب هستند. برای مثال داده های مورد استفاده در اینجا شامل نظرات کاربران می باشد در این نظرات محتوا و اندازه متن متفاوت می باشد و شامل دسته ای از کلمات مانند این، آن، و، یا، ای، موجی ها و... می باشد که در پردازش آن تغییر محسوسی در نتیجه خروجی نمی دهد و فقط زمان پردازش را افزایش می دهد. پس نیاز است که قبل از پردازش داده ها تمیز شوند. این دسته از کلمات به عنوان کلمات وقفه در فارسی (Stop words in Persian) شناخته می شوند که قبل از هر پردازشی لازم که این کلمات از متن و نظرات حذف شود تا پردازش متن دقیق تر و سریع تر صورت گیرد. در این مرحله داده های گم شده هم از داده ها حذف می شوند. برای مثال نظر کاربر حذف شده است ولی برچسب آن می باشد اگر این نوع داده وارد شبکه شود ماشین داده خالی را با برچسب آن یاد می گیرد و باعث کاهش دقت خروجی ماشین می شود. در مرحله تمیز کردن داده ها پروسه ی دیگری که انجام میشود برداری کردن داده ها می باشد در حقیقت داده ها (کلمه ها) به اعداد قابل فهمی برای ماشین تبدیل می شوند به این پروسه Vectorization گفته می شود.

## ۳-۳-۴- تقسیم داده ها (Split data)

در این مرحله داده های برداری مرحله قبل به دو دسته تقسیم می شوند

۱. داده های آموزش (Train Data)

## ۲. داده های تست (Test Data)



شکل ۸ تقسیم داده ها

## ۴-۳-۴- داده های آموزش (Train Data)

در این مرحله قسمت ای از داده های مرحله قبل، که غالبا ۷۰ درصد داده ها می باشد، وارد مدل یادگیری ماشین تعیین شده می شود و مدل ما توسط داده ها آموزش داده می شود. در حقیقت قسمت اصلی یادگیری ماشین، قسمت یادگیری می باشد که در این مرحله صورت میگیرد. به عبارت دیگر مدل ما توسط داده ها تغذیه می شود.



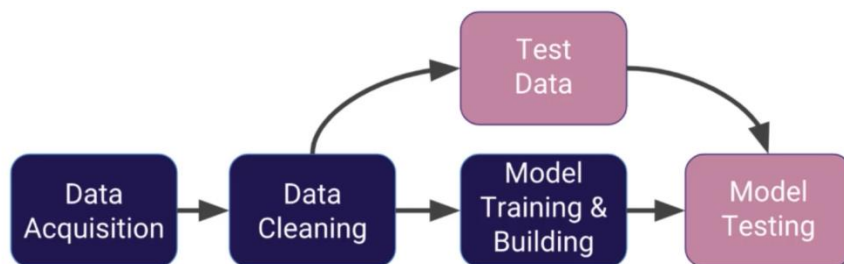
شکل ۹ آموزش مدل

## ۴-۳-۵- داده های تست (Test Data)

پس از آموزش دادن مدل ما نیاز است که مدل ما ارزشیابی شود که متوجه دقت یادگیری و میزان خطای مدل خود شویم. برای انجام این موضوع لازم است که مدل ما تست شود.

برای تست کردن مدل لازم است داده هایی که مدل از قبل ندیده است، که غالباً ۳۰ درصد از داده های مرحله قبل می باشد، را به عنوان ورودی به مدل دهیم و خروجی مدل را مشاهده کنیم و میزان دقت و خطای مدل را اندازه گیری کنیم. توجه شود که برای عدم تقلب در این موضوع دادهایی باید به شبکه داد که شبکه قبلاً آن را ندیده باشد.

در قسمت ارزیابی ماتریس های ارزیابی (Evaluation Metrics) معرفی می شود که در ادامه فصل توضیحات بیشتر داده می شود.



شکل ۱۰ تست کردن مدل

در مثالی که در اینجا مورد بحث می باشد داده های ما شامل دسته ای از نظرات و دسته ای از برچسب ها می باشد ورودی ها که نظرات باشند پیشوند  $X$  و خروجی که برچسب ها می باشد پیشوند  $Y$  را میگیرند. در کل چهار دسته داده داریم

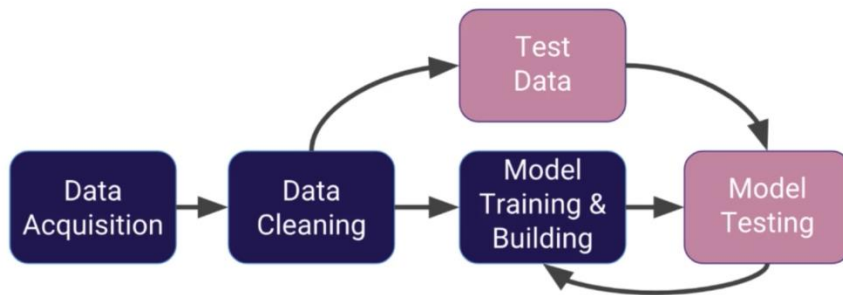
$X_{train}$ ,  $Y_{train}$ ,  $X_{test}$ ,  $Y_{test}$

#### ۴-۳-۶- بهبود مدل (Improve Model)

پس از ارزیابی مدل توسط داده های تست الگوریتم طراحی شده است که باعث بهبود یادگیری ماشین می شود در حقیقت پس از ارزیابی مدل مرحله یادگیری مجدداً انجام می شود تا زمانی که به حداکثر دقت قابل دستیابی رسید در این مرحله یادگیری متوقف می شود.

به زبان دیگر پس از ارزیابی مدل توسط داده های تست میتوان با ایجاد تغییر در پارامتر های یادگیری، یادگیری را بهبود بخشید.

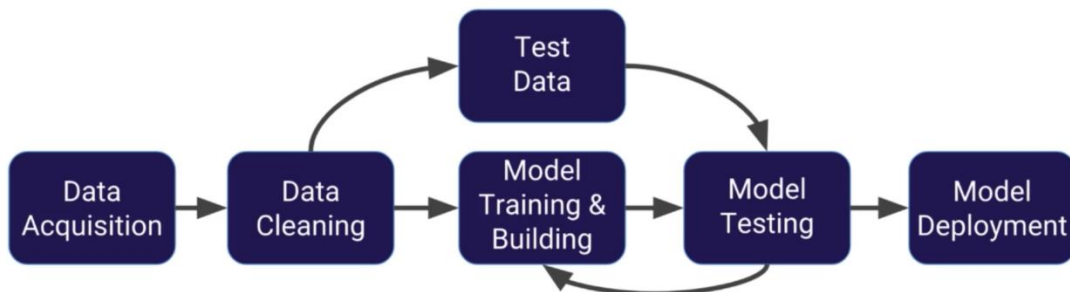




شکل ۱۱ بهبود مدل

### ۴-۳-۷- گسترش مدل (Model Deployment)

در مرحله آخر مدل آماده استفاده و گسترش می باشد.



شکل ۱۲ گسترش مدل

### معیارهای طبقه بندی (Classification Metrics)

در قسمت قبل اشاره شد که برای ارزیابی مدل ، داده هایی که مدل از قبل ندیده باشد به عنوان ورودی وارد مدل می شود و خروجی آن مشاهده می شود و میزان دقت و خطای مدل بررسی می شود. برای ارزیابی بهتر از دسته ماتریس های Classification Metrics استفاده می شود. که در این قسمت قرار به صورت کامل و دقیق مورد بررسی قرار گیرد. در حقیقت فقط بررسی میزان صحت(دقت) برای ارزیابی مدل کافی نمی باشد و نیاز است پارامترهای دیگری وارد شود.

#### ۴-۴- معیارهای طبقه بندی (Classification Metrics)

به صورت کامل classification metrics شامل پارامترهای زیر می باشد:

- Accuracy
- Recall
- Precision
- F1-Score

توجه شود که چهار ویژگی بیان شده از کلمات انگلیسی استفاده می شود تا درک آن راحتتر باشد و کمتر باعث سردگمی در فهم موضوع شود چرا که برای مثال مورد اول و سوم در فارسی معادل هم هستند و بیان فارسی آن باعث ایجاد سردگمی می شود.

برای فهم بهتر موضوع و درک کاربرد این موضوع ادامه ی مبحث را با مثالی واقعی جلو میبریم تا فهم موضوع را آسان تر کند.

در داده های ما در پیشبینی خروجی دو نتیجه دارد یا خروجی درست (Correct) می باشد یا خروجی اشتباه (Incorrect) می باشد.

با توجه به داده های ما خروجی می تواند دو حالت داشته باشد که همان دو حالت برچسب شده ی داده های ما می باشد

۱. خرید کالا توسط کاربر پیشنهاد می شود

۲. خرید کالا توسط کاربر پیشنهاد نمی شود

که این نوع دسته بندی به نام binary classification شناخته میشوند.

#### Accuracy

این پارامتر برای اندازه گیری میزان دقتی که مدل ما می تواند پیشبینی کند می باشد برای بدست آوردن این مقدار کافی است که تعداد پیش بینی های صحیح را به کل پیشبینی ها تقسیم کنیم

$$\text{accuracy} = \frac{\text{تعداد پیشبینی درست}}{\text{تعداد کل پیشبینی ها}} * 100$$

برای مثال اگر در خروجی تعداد پیش بینی درست ۹۰ تا باشد و تعداد کل پیشبینی ها ۱۰۰ باشد دقت مدل ما ۹۰ درصد می باشد.

Accuracy برای بررسی مدلی قابل قبول است که داده ها در حالت متعادل باشد به این معنی که تعداد داده ها در دو حالت باینری تعداد نزدیک به هم داشته باشند. برای مثال در داده های ما تعداد نظرات که خرید کالا را پیشنهاد می کنند با تعداد کاربرانی که خرید کالا را پیشنهاد نمی کنند تعداد نزدیک به همی دارد. حال فرض شود اگر ۹۹ تا نظر مثبت داشته باشیم و ۱ نظر منفی اگر مدل ما برای همه ی ورودی ها خروجی مثبت را پیشبینی کند accuracy مدل ما همیشه ۹۹ درصد می باشد که این روش مناسبی برای ارزیابی مدل ما نمی باشد. در حقیقت این پارامتر برای داده های متعادل مناسب نمی باشد. به این منظور پارامتر دیگر Precision, Recall پیشنهاد میشود که در ادامه توضیح داده می شود. برای متوجه شدن پارامتر های دیگر در ابتدا نیاز است که قبل از آن مبحث دیگری به اسم Confusion metrics بیان شود که در ادامه توضیحات لازم داده می شود.

#### ۴-۵- ماتریس پریشانی (Confusion Matrix)

همانطور که در قسمت قبل اشاره شد برای متوجه شدن و به دست آوردن مقدار Precision و Recall لازم است ماتریس پریشانی تعریف و توضیح داده شود. به دلیل اینکه معنای بعضی کلمات انگلیسی در فارسی ترجمه سلیسی ندارند از اصل کلمات استفاده شده است ولی هر قسمت توضیحات لازم برای فهم بهتر آن داده شده است. فرم کلی ماتریس پریشانی به صورت زیر می باشد:

		predicted condition	
total population		prediction positive	prediction negative
true condition	condition positive	<b>True Positive (TP)</b>	<b>False Negative (FN)</b> (type II error)
	condition negative	<b>False Positive (FP)</b> (Type I error)	<b>True Negative (TN)</b>

شکل ۱۳ ماتریس پریشانی

برای راحت درک کردن مفهوم بالا همراه با مثالی واقعی موضوع را پیش میبریم باتوجه به داده های ما خروجی دو حالت دارد:

۱. پیشنهاد می شود

۲. پیشنهاد نمی شود

برای راحتی موضوع حالت اول را خروجی مثبت و حالت دوم را خروجی منفی در نظر میگیریم مدل آموزش دیده برای داده های جدید خروجی را پیشبینی (Predict) می کند این خروجی می تواند مثبت یا منفی باشد دو ستون ماتریس پریشانی را این دو حالت تشکیل می دهد. برای هر خروجی مقدار صحیح آن هم وجود دارد که این حالت هم میتواند دو وضعیت داشته باشد وضعیت اول مثبت و وضعیت دوم منفی می باشد توجه شود این دو وضعیت، وضعیت صحیح و خروجی مورد انتظار ما می باشد .

حال با توجه به خروجی پیشبینی شده و وضعیت مورد انتظار ما چهار حالت تعریف می شود.

۱. True Positive (TP)

۲. False Positive (FP) - Error type I

۳. False Negative (FN) - Error type II

۴. True Negative (TN)

هر چهار حالت در ادامه توضیح داده می شود.

### **True Positive**

این حالتی متعلق به زمانی است که خروجی مثبت پیش بینی می شود و وضعیت مورد انتظار هم مثبت بوده است این خروجی مطلوب است. برای مثال مدل ما نظر مثبتی را مثبت پیش بینی میکند.

### **Error type I-(FP)False Positive**

این حالت زمانی است که مدل ما خروجی مثبت را پیش بینی می کند ولی خروجی مطلوب ما منفی می باشد. برای مثال مدل ما نظر منفی را مثبت پیش بینی کرده است. در این حالت خطا رخ داده است و دچار خطای نوع اول شده است.

### **Error type II-(FN)False Negative**

در این حالت خروجی مدل منفی پیش بینی شده است در حالی که خروجی مورد انتظار مثبت بوده است برای مثال مدل نظر مثبتی را منفی پیش بینی کرده است. در این حالت خطای نوع دوم اتفاق افتاده است.

### **(TN)True Negative**

این حالت مانند حالت اول خروجی مطلوب رخ داده است یعنی مدل ما خروجی منفی را پیش بینی کرده است و خروجی مطلوب ما هم نیز منفی بوده است.

حال با توجه به ماتریس پریشانی دو حالت Precision و Recall تعریف می شود.

### **Recall**

Recall در حالتی مورد بررسی قرار میگیرد که خروجی مطلوب ما مثبت می باشد به عبارت دیگر وضعیت خروجی مثبت است؛ حال نسبت خروجی پیش بینی شده مثبت در وضعیت مثبت به کل تعداد پیش بینی در وضعیت مثبت مقدار Recall را در اختیار ما قرار می دهد.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} * 100$$

### Precision

Precision در حالتی مورد بررسی قرار میگیرد که خروجی پیش بینی ما مثبت می باشد ؛ و بررسی میکند چه تعداد از پیش بینی مثبت مدل درست پیش بینی شده است .

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} * 100$$

### F1-Score

F1-Score رابطه ای است بین Precision و Recall می باشد. در حقیقت میانگین هارمونیک Precision و Recall می باشد. علت استفاده از این پارامتر به جای استفاده از میانگین این است که مقادیر عجیب را نمایش دهد.

برای مثال اگر مقدار precision برابر ۱ باشد و Recall برابر صفر شود مقدار میانگین برابر ۰.۵ می باشد ولی مقدار F1-Score برابر صفر می باشد که نتیجه قابل توجه تری می باشد.

$$\text{Precision} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Accuracy در قبل تعریف شد ولی میتوان آن را هم نیز از ماتریس پیرایشانی استفاده کرد که فرمول آن به صورت زیر است .

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} * 100$$

برای درک بهتر مثالی عددی را در ادامه بررسی می کنیم. فرض کنید ماتریس پیرشانی زیر در اختیار ما می باشد .  
کل ورودی ۱۶۵ می باشد.

n = 165	Predicted: No	Predicted: Yes
Actual: No	50	10
	5	100
Actual: Yes		

شکل ۱۴ مثال ماتریس پیرشانی

حال چهار مقدار Accuracy, Recall, precision, F1-

### Accuracy

$$\text{Accuracy} = \frac{100 + 50}{100 + 5 + 10 + 50} * 100 = 90$$

### Recall

$$\text{Recall} = \frac{100}{100 + 5} * 100 = 95$$

## Precision

$$\text{Precision} = \frac{100}{100 + 10} * 100 = 91$$

## F1-Score

$$\text{F1 - Score} = \frac{2 * 95 * 91}{91 + 95} = 92$$

در قسمت اولیه این فصل اشاره به موضوع Vectorization شد در ادامه توضیح کامل تری در این مورد می دهیم.

## ۴-۶- استخراج ویژگی‌ها (Feature Extraction)

استخراج متن در حقیقت استخراج اطلاعات مهم یک متن می باشد هدف این موضوع استخراج اطلاعات مهم متن و پردازش دقیق تر در یادگیری ماشین می باشد. در تحلیل متن ها که شامل کلمات زیادی می باشد دسته ای از کلمات وجود دارند در تحلیل اهمیت بالایی دارند و دسته ای از کلمات هستند که در نتیجه ی تحلیل موثر نمی باشند در حقیقت اهمیت کلمات در هر متنی متفاوت می باشد. با توجه به این موضوع اهمیت استخراج ویژگی ها نمایان می شود.

روش های متعددی برای استخراج ویژگی ها وجود دارد که در ادامه توضیح داده خواهد شد.



### ۱-۶-۴- بردار سازی Vectorization

ماشین ها قابلیت درک کلمات را ندارند و کامپیوتر ها فقط اعداد را میتوانند درک کنند. در پردازش طبیعی به روش یادگیری ماشین با نظارت نیاز است که در مرحله ی قبل از یادگیری داده های ما به دسته ای از اعداد تبدیل شود تا ماشین توانایی درک و یادگیری را داشته باشد به این پروسه ی تبدیل متن به اعداد Vectorization گفته می شود.

به کمک Vectorization کردن متن میتوانیم ویژگی های متن را استخراج کنیم و پردازش های لازم را انجام دهیم در حقیقت با این عمل Feature Extraction یا همان استخراج ویژگی ها را انجام داده ایم.

روش های متعددی برای بردار سازی متن وجود دارد. روشی که در این جا استفاده شده است به کمک دسته ای از مفاهیم کمکی که در ادامه توضیح داده خواهد استفاده شده است.

### ۲-۶-۴- Count Vectorization

در ابتدای این پروسه هر متن به کلمات تشکیل دهنده ی آن تبدیل می شود سپس با توجه به تکرار هر کلمه در متن عددی برای آن عدد اختصاص داده می شود. برای مثال:

متن اول: "از خرید این کتاب راضی هستم."

متن دوم: "خرید این گوشی را پیشنهاد میکنم. گوشی خوب است"

حال برای هر متن به صورت جداگانه برای هر کلمه ی متن یک عدد اختصاص داده می شود که آن عدد برابر تعداد تکرار آن کلمه در متن می باشد.

کلمه	از	خرید	این	کتاب	راضی	هستم	گوشی	را	پیشنهاد	میکنم	خوبی	است
متن اول	۱	۱	۱	۱	۱	۱	۰	۰	۰	۰	۰	۰
متن دوم	۰	۱	۱	۰	۰	۰	۲	۱	۱	۱	۱	۱

جدول ۱ Count Vectorization

Count Vectorization نمیتواند به تنهایی حاوی اطلاعات مهمی باشد چرا که بعضی از کلمات دارای تکرار زیاد می باشند مانند این آن و یا ... در حالی که تکرار زیاد این کلمات برای پردازش مفید نمی باشند و حاوی اطلاعات مهمی نمی باشند بنابر نیاز است از روش کامل تری استفاده شود روشی که برای بهبود این مشکل بیان می شود TF-IDF Vectorizer می باشد.

### ۳-۶-۴ - TF-IDF Vectorizer

TF-IDF شامل دو قسمت می باشد

۱. Term Frequency (TF)

۲. Inverse Document Frequency (IDF)

#### Term Frequency (TF)

قسمت اول برای هر کلمه از متن یک عدد اختصاص داده می شود که برابر تعداد تکرار آن کلمه در همان متن می باشد دقیقاً مانند قسمت Count Vectorizer؛ ولی همانطور که اشاره شد این حالت دارای مشکل می باشد چرا که برای بعضی از کلمات که تکرار زیاد دارند مانند این، آن، و، یا... عدد بزرگی می شود در حالی این کلمات در پردازش متن از اهمیت کمی برخوردار می باشد برای حل این مشکل قسمت دوم (IDF) بیان می شود.

#### Inverse Document Frequency (IDF)

در این قسمت هر متن یا به صورت خاص در این مقاله، هر نظر به عنوان یک سند (Document) در نظر گرفته می شود. برای هر کلمه علاوه بر این که تعداد تکرار آن کلمه در آن متن عددی در نظر گرفته می شود عددی دیگر هم اختصاص داده می شود؛ به این صورت که برای هر کلمه بررسی می شود که به صورت فرمول زیر محاسبه می شود.

$$idf(w, D) = \log \frac{N}{|\{w \in D : w \in d\}|}$$

هر کلمه:  $w$

تعداد کل سند ها:  $N$

کل سند ها:  $D$

هر سند:  $d$

اختصاص این عدد برای هر کلمه به این گونه است که بررسی می شود هر کلمه در چه تعداد سند تکرار شده است و مقدار لوگاریتم نسبت تعداد کل سند ها به تعداد سندی که آن کلمه در آن تکرار شده است به آن کلمه اختصاص داده می شود.

برای فهم بهتر موضوع مثال زیر را بررسی میکنیم

۱. "من از خرید این کتاب راضی هستم"

۲. "من از خرید این لپ تاپ ناراضی هستم"

کلمه	من	از	خرید	این	کتاب	راضی	هستم	لپ تاپ	ناراضی
متن اول	۱	۱	۱	۱	۱	۱	۱	۰	۰
متن دوم	۱	۱	۱	۱	۰	۰	۱	۱	۱
$ \{w \in D : w \in d\} $	۲	۲	۲	۲	۱	۱	۲	۱	۱
$\log \frac{N}{ \{w \in D : w \in d\} }$	۰	۰	۰	۰	۰.۳	۰.۳	۰	۰.۳	۰.۳

جدول TF-IDF<sup>۲</sup>

همانطور که مشاهده میکنید هر نظر به عنوان یک سند مجزا در نظر گرفته شده است و برای هر کلمه دو مقدار در نظر گرفته شده است مقدار اول که تعداد تکرار هر کلمه می باشد به تنهایی مفید نمی باشد ولی با اختصاص دادن مقدار  $idf$  برای هر کلمه ارزش هر کلمه مشخص تر می باشد برای مثال کلمه من، از، خرید، این، هستم در هر دو سند تکرار شده است ولی مقدار  $idf$  صفر بدست آمده است و این مشخص میکند که این دسته از کلمات با اینکه دارای تکرار زیاد می باشند ولی حاوی اطلاعات مفیدی نیستند بلکه فقط در همه ی سند ها تکرار شده اند اما کلمات کتاب، لپ تاپ، راضی، ناراضی دارای

مقدار غیر از صفر می باشند و حاوی اطلاعات مهم می باشند و در پردازش متن می توانند مفید واقع شوند.

نکات ابتدایی و مورد استفاده برای این فصل بیان شد حال به کمک کتابخانه های در دسترس مراحل بالا را برای داده های دیجیکالا در محیط پایتون عملی میکنم و نتایج خروجی را مشاهده میکنم. توجه شود در این مقاله آموزش پایتون صورت نمیگیرد ولی تاجایی که مقدور باشد کدها را به صورت خط به خط توضیح میدهیم .

#### ۷-۴- آشنایی با Sckit-learn در Python و اعمال آن بر داده های واقعی

Sckit-learn کتابخانه ای در دسترس برای پایتون که شرایط تحلیل داده ها را برای ما فراهم می کند.

اضافه کردن کتابخانه ها:

```
>>>import numpy as np
>>>import pandas as pd
>>>from sklearn.model_selection import train_test_split
>>>from sklearn.pipeline import Pipeline
>>>from sklearn.feature_extraction.text import TfidfVectorizer
>>>from sklearn.svm import LinearSVC
>>>from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
```

اضافه کردن داده ها :

```
>>>df=pd.read_excel(r'digi.xlsx',sep='\t')
>>>df.dropna(inplace=True)
>>>ind=df[df['recommend']== 'no_idea'].index.tolist()
>>>df.drop(ind,inplace=True)
>>>ind1=df[df['recommend']== 're'].index.tolist()
>>>df.drop(ind1,inplace=True)
>>>X=df['comment']
>>>y=df['recommend']
>>>X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=42)
>>>text_clf=Pipeline([('tfidf',TfidfVectorizer()),('clf',LinearSVC())])
>>>text_clf.fit(X_train,y_train)
>>>predictions=text_clf.predict(X_test)
>>>print("Confusion Matrix\n",confusion_matrix(predictions,y_test))
>>>print("Classification Report\n",classification_report(predictions,y_test))
>>>print("Accuracy\n",accuracy_score(predictions,y_test))
```

ورود نظر جدید به شبکه :

```
>>> pre=text_clf.predict(['باتریش زود شارژ خالی میکنه و خیلی سنگینه و اصلا خوش فرم نیست'])
```

خروجی:

```
print("result\n",pre)
if pre=='not_recommended':
    print('پیشنهاد نمی شود')
else:
    print('پیشنهاد می شود')
```

برای مثال بالا چیزی که در خروجی مشاهده می کنیم به شرح عکس زیر می باشد:

```

/home/liamirpy/Desktop/NLTK/venv/bin/python /home/liamirpy/Desktop/NLTK/dgk-classification.py
['not_recommended' 'recommended' 'recommended' ... 'recommended'
 'recommended' 'recommended']
Confusion Matrix
[[ 973 195]
 [ 307 3209]]
Classification Report
              precision    recall  f1-score   support

not_recommended    0.76    0.83    0.79    1168
recommended        0.94    0.91    0.93    3516

accuracy          0.89    0.89    0.89    4684
macro avg         0.85    0.87    0.86    4684
weighted avg      0.90    0.89    0.89    4684

Accuracy
0.8928266438941076
result
['not_recommended']
پیشنهاد نمی شود

Process finished with exit code 0

```

شکل ۱۵ خروجی طبقه بندی متن

همانطور که در شکل میبینید خروجی شامل چند قسمت می باشد

- Confusion Matrix
- Classification Report
- Accuracy
- result

دقت این مدل آموزش دیده تقریباً ۸۹ درصد می باشد که دقت قابل قبولی می باشد و خروجی مشاهده شده برای متن زیر

- "باتریش زود شارژ خالی میکنه و خیلی سنگینه و اصلاً خوش فرم نیست"

منفی یا پیشنهاد نمی شود می باشد. در حقیقت مدل ما پیش بینی می کند که این کاربر با این نظر خرید کالا را پیشنهاد نمی کند. که به نظر خروجی معقولی می باشد.

در این فصل تحلیل از روش طبقه بندی متن را بررسی کردیم و مدلی طراحی کردیم و خروجی را برای مثالی مشاهده کردیم و نتیجه مطلوبی مشاهده شد.  
در فصل بعدی از روش دیگر، تحلیل احساسات برای طراحی مدل استفاده می شود.

## فصل پنجم

### تحلیل احساسات

در فصل قبل روش دسته بندی متن (Text Classification) استفاده شد که متن را به دسته ای از اعداد تبدیل کردیم و پردازش را انجام دادیم در حقیقت مدل خود را به کمک دسته ای از اعداد آموزش دادیم روشی دیگر که برای پردازش زبان طبیعی بیان می شود تحلیل احساسات یا Sentiment Analysis می باشد که در ادامه مورد بحث قرار داده می شود.

#### ۱-۵- تحلیل احساسات (Sentiment Analysis)

به زبان ساده هدف تحلیل احساسات بررسی میزان حس یک متن می باشد در حقیقت این روش از پردازش به دنبال این است که متوجه شود یک جمله یا یک متن مثبت است یا منفی اگر بخواهیم دقیق تر بیان کنیم این روش پردازش بررسی میکند چه میزان یک متن مثبت و چه میزان منفی است. برای مثال جمله به دو جمله زیر توجه کنید :

۱. "من خیلی ناراضیم از خریدم و اصلاً کیفیتش خوب نیست"

۲. "جنسش خیلی خوبه و خوش فرم و خوش دسته"

وقتی شما جمله اول میخوانید حس نویسنده ی متن که به شما القا می شود حس منفی و نارضایتی می باشد در حالی که جمله دوم حس مثبت و رضایت را منتقل می کند. روشی که مغز شما متوجه این موضوع می شود این است که شما از قبل مثبت یا منفی بودن کلمات را آموخته اید برای مثال شما از



قبل آموزش دیده اید که کلمه ی ناراضیم دارای بار معنایی منفی و کلمه خوش فرم بار معنایی مثبت می باشد سپس شما مجموع بار معنایی هر جمله را به کمک بار معنایی کلمات به صورت جداگانه بررسی میکنید و بار معنای کلی جمله را متوجه می شوید .

در آموزش ماشین هم نیز میتوان از این روش استفاده کرد به این صورت که ماشین را نسبت به بار معنایی هر کلمه آموزش می دهیم .روشی که استفاده می شود vader نام دارد.

## ۲-۵- (VADER)Valance Aware Dictionary for sEntiment reasoning

VADER یک مدلی می باشد که برای تحلیل احساسی مورد استفاده قرار میگیرد که به دو قطب مثبت و منفی و همچنین به شدت احساس ، حساس می باشد. که در پکیج NLTK در پایتون در دسترس می باشد.

در این مدل دیکشنری از کلمات با شدت حالت (احساس) هر کلمه وجود دارد این شدت احساس با بازه ای از اعداد تعریف شده است در واقع برای هر کلمه عددی در حکم شدت احساس آن کلمه در نظر گرفته شده است .

در تحلیل احساسات مجموع شدت این کلمات در یک جمله یا متن ، میزان منفی یا مثبت جمله یا متن را نشان می دهد.

اگر توجه کرده باشد در فصل قبل داده هایی که برای آموزش مدل استفاده شده، داده های برچسب شده (Labeled) بود و از روش یادگیری با نظارت استفاده شد (Supervised Learning) حال اگر داده های در دسترس داده هایی باشد که برچسب نشده باشد چگونه میتوان مدل را آموزش داد؟

در داده های در اختیار قرار داده شده ی دیجیکالا نظرات برچسب زده شده اند ولی ما در اینجا قصد داریم با فرض برچسب نشدن داده ها چگونه می توان پردازش زبان طبیعی را انجام دهیم.

برای حل این مشکل نیاز است که از مدلی استفاده کنیم که قابلیت برچسب کردن داده ها را داشته باشد ولی نیازی به نیروی انسانی نباشد چرا که حجم داده ها زیاد می باشد .

روشی که برای حل این مشکل پیشنهاد شده است تحلیل احساسات هر نظر می باشد و با توجه به شدت احساس هر نظر اگر مجموع شدت تمام کلمات هر نظر منفی شد برچسب "پیشنهاد نمی شود" و اگر مجموع شدت تمام کلمات هر نظر مثبت شد برچسب "پیشنهاد می شود" برای آن نظر در گرفته شود. متأسفانه مشکلی که وجود دارد استفاده از پکیج VADER فقط برای دسته ای زبان های مانند انگلیسی، فرانسوی وجود دارد و برای فارسی هنوز این پکیج وجود ندارد.

برای حل این مشکل راه حلی که به ذهن من رسید به این صورت می باشد که به کمک کتابخانه ای در پایتون می توان به مترجم گوگل (google translate) متصل شد و متن را به زبان دیگر ترجمه کرد. در این مقاله ما به کمک این کتابخانه در ابتدا تمام نظرات را به انگلیسی ترجمه کردیم و سپس تحلیل احساسات را برای هر نظر انجام داده ایم و به این روش هر نظر را برچسب زده ایم. در مرحله آخر برای بدست آوردن میزان خطا برچسب های ایجاد شده به کمک این روش را با برچسب های اصلی مقایسه می کنیم و میزان خطا را بدست می آوریم .

### ۳-۵- کدنویسی تحلیل احساسات در پایتون:

کد نویسی این قسمت دارای دو مرحله می باشد :

#### ۱-۳-۵- مرحله اول برنامه نویسی:

در مرحله اول برنامه نویسی تمام نظرات کاربران به انگلیسی ترجمه می شود:

```
>>>from googletrans import Translator
>>>import pandas as pd
>>>import numpy as np
>>>import string
>>>import csv
>>>import re
>>>df=pd.read_excel(r'digi.xlsx',sep='\t')
>>>df.dropna(inplace=True)
>>>ind=df[df['recommend'] == 'no_idea'].index.tolist()
>>>df.drop(ind,inplace=True)
>>>ind1=df[df['recommend'] == 're'].index.tolist()
>>>df.drop(ind1,inplace=True)
>>>comment=df['comment'].to_csv('comment.csv')
>>>comment = pd.read_csv('comment.csv', header=None)
>>>comment.rename(columns={0: 'num', 1: 'comment'}, inplace=True)
>>>comment=comment.to_csv('comment.csv', index=False)
>>>label=df['recommend'].to_csv('label.csv')
>>>label = pd.read_csv('label.csv', header=None)
>>>label.rename(columns={0: 'num', 1: 'label'}, inplace=True)
>>>label=label.to_csv('label.csv', index=False)
>>>comment = pd.read_csv('comment.csv')
>>>comment_label=comment.merge(label)
>>>comment_label=comment_label.to_csv('comment-label.csv', index=False)
>>>comment=pd.read_csv('comment.csv')
>>>comment.dropna(inplace=True)
```

```
>>>com=[]
for k in range(len(comment)):
    com.append(comment.values[k][1])
```

تمیز کردن داده ها :

```
>>>for u in range(len(com)):
com[u]=com[u].replace('._x000D_', ' ')
com[u]=com[u].replace('._x000D_', ' ')
com[u]=com[u].replace(' ', '😊')
com[u]=com[u].replace(' ', '😄')
com[u]=com[u].replace(' ', '😐')
com[u]=com[u].replace(' ', '😞')
com[u]=com[u].replace(' ', '😍')
com[u]=com[u].replace(' ', '😎')
com[u]=com[u].replace(' ', '😘')
com[u]=com[u].replace(' ', '😏')
com[u]=com[u].replace(' ', '😂')
com[u]=com[u].replace(' ', '❤️')
com[u]=com[u].replace(' ', '😭')
com[u] = com[u].replace('\n', ' ')
>>>tr=np.load('my_file.npy', allow_pickle=True).item()
>>>translator = Translator()
>>>for i in range(0,len(com)):
t = translator.translate(com[i])
tr[com[i]]=str(t.text)
np.save('my_file.npy',tr)
print(i)
read_dictionary = np.load('my_file.npy', allow_pickle=True).item ()
print(len(read_dictionary))
>>>read_dictionary = np.load('my_file.npy', allow_pickle=True).item()
>>>with open('comment-translated.csv','w') as f:
w = csv.writer(f)
row = ['comment', 'translate']
w.writerow(row)
w.writerows(read_dictionary.items())
f.close()
>>>csv_test= pd.read_csv('comment-translated.csv')
```

```
>>>comment_label= pd.read_csv('comment-label.csv')
>>>translate_text_label=csv_test.merge(comment_label,on='comment')
>>>translate_text_label=translate_text_label.to_csv('comment-translate-label.csv', index=False)
```

در مرحله دوم برنامه شبکه را آموزش می دهیم و داده جدید را به شبکه می دهیم و خروجی را مشاهده می کنیم:

```
>>>import nltk
>>>from googletrans import Translator
>>>from nltk.sentiment.vader import SentimentIntensityAnalyzer
>>>import pandas as pd
>>>from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
>>>nltk.download('vader_lexicon')
>>>sid=SentimentIntensityAnalyzer()
>>>df=pd.read_csv('comment-translate-label.csv')
>>>df.dropna(inplace=True)
>>>blanks=[]
>>>df['scores']=df['translate'].apply(lambda re:sid.polarity_scores(re))
>>>df['compound']=df['scores'].apply(lambda com:com['compound'])
>>>df['compound_score']=df['compound'].apply(lambda score:'pos' if score>=0 else 'neg')
>>>df['label_score']=df['label'].apply(lambda score:'pos' if score=='recommended' else 'neg')
>>>score=df.to_csv('score.csv', index=False)
>>>print(accuracy_score(df['label_score'],df['compound_score']))
>>>w='باطریش یکم ضعیفه ولی در کل کیفیت تصویر و صداش خوبه ارزش خرید داره'
>>>translator = Translator()
>>>t = translator.translate(w)
>>>result=sid.polarity_scores(t.text)
>>>print(result)
>>>print(result['compound'])
>>>if result['compound']<0:
    print('پیشنهاد نمی شود')
else:
    print('پیشنهاد می شود')
```

برای کد بالا خروجی که مشاهده می شود به صورت زیر می باشد:

```

/home/liamirpy/Desktop/NLTK/venv/bin/python /home/liamirpy/Desktop/NLTK/digikala-sentiment.py
/home/liamirpy/Desktop/NLTK/venv/lib/python3.7/site-packages/nltk/twitter/__init__.py:20: UserWarning: The twython
warnings.warn("The twython library has not been installed. ")
[nltk_data] Downloading package vader_lexicon to
[nltk_data] /home/liamirpy/nltk_data...
[nltk_data] Package vader_lexicon is already up-to-date!
0.8162284678150499
پیشنهاد می شود

Process finished with exit code 0
  
```

#### شکل ۱۶ خروجی تحلیل احساسات

همانطور که مشاهده می کنید خروجی دارای دقت ۸۱ می باشد به این معنا که با تحلیل احساسات توانسته ایم ۸۱ درصد بر چسب داد های تست را درست پیش بینی نماییم .

و برای ورودی متن زیر :

- 'باطریش یکم ضعیفه ولی در کل کیفیت تصویر و صداش خوبه ارزش خرید داره'

خروجی مثبت یا پیشنهاد می شود را پیش بینی کرده است که خروجی مطلوب می باشد.

در این فصل تحلیل از روش تحلیل احساسات را بررسی کردیم و نتیجه ی مطلوب را مشاهده کردیم ولی دقت این حالت ۸۱ درصد می باشد در حالی که دقت در تحلیل طبقه بندی متن ۸۹ درصد می باشد ولی با این حال این نوع تحلیل باز هم مورد قبول می باشد چرا که همیشه داده های ما برچسب شده نمی باشند.

## فصل شیشم

### مدل سازی موضوع

در دو فصل گذشته تحلیل های ابتدایی را انجام دادیم و از طریق دو روش دو مدلی آموزش دیده ای را ایجاد کردیم که آماده دریافت داده های جدید و پیش بینی خروجی می باشد. ورودی هر مدل نظر جدیدی می باشد که خروجی پیش بینی می کند این کاربر با ایجاد این نظر خرید کالا را پیشنهاد می کند یا پیشنهاد نمی کند.

در ادامه قصد داریم مدل جدیدی را ایجاد کنیم که قابلیت این را داشته باشد که نظر ایجاد شده توسط کاربر در صورت مثبت بودن نظر چه ویژگی از کالا از نگاه کاربر مثبت و اگر نظر ایجاد شده توسط کاربر منفی می باشد چه ویژگی از کالا از نگاه کاربر منفی می باشد.

برای ایجاد این مدل روشی که پیشنهاد می شود روش مدل سازی موضوع (Topic modeling) استفاده شده است.

#### ۱-۶- مدل سازی موضوع (Topic Modeling)

مدل سازی موضوع زمانی نیاز می شود که حجم داده های زیادی در اختیار است و قصد داریم اسناد را دسته بندی کنیم؛ برای مثال اسناد یک روزنامه را در اختیار داریم و قصد داریم هر کدام از اسناد را موضوع بندی کنیم ساده ترین کار مطالعه کامل متن و اختصاص دادن موضوعی (Topic) برای آن متن است ولی واضح است که در حجم بالای داده ها این کار مقدور نمی باشد.

نکته قابل توجهی که وجود دارد این است که ما فقط داده هایی در اختیار داریم که دارای برچسب نمی باشند (Unlabeled) و به طبع برای ساختن مدل باید از یادگیری بدون نظارت (Unsupervised Learning) استفاده کرد ولی تفاوت مهمی که با یادگیری بدون نظارت فصل قبل دارد این است که

نمی تواند خطایی را اندازه گیری کنیم چرا که جواب درستی در اختیار نداریم که نسبت به آن خطا را اندازه گیری و میزان صحت مدل را اندازه گیری کنیم. روشی که برای یادگیری بدون نظارت در این زمینه استفاده می شود روش خوشه بندی (Clustering) می باشد.

## ۲-۶- Latent Dirichlet Allocation (LDA)

دریکلت ریاضی دان آلمانی که در سال ۱۸۰۰ میلادی در زمینه ی ریاضی مدرن فعالیت داشت که پس از وی توزیع احتمالی به نام توزیع دریکلت (Dirichlet Distribution) تعریف شد که LDA بر اساس این توزیع بیان شد که در سال ۲۰۰۳ مقاله ای در مورد پیدا کردن موضوع (Topic discovery) نوشته شد.

روش LDA برای پیدا کردن موضوع بر دو فرض استوار است:

- اسنادی که دارای موضوع یکسان می باشند از دسته ای از کلمات شبیه به هم استفاده میکنند و برای هر دسته از کلمات موضوعی را در نظر میگیریم برای مثال متونی که در مورد اقتصاد است از دسته ای از کلمات شبیه به هم و مرتبط با اقتصاد استفاده می کند
- با دسته بندی کردن گروهی از کلمات مرتبط می تواند متون جدید را باتوجه به میزان تکرار هر یک از دسته کلمات، موضوع بندی کرد.

در ابتدای کار تعدادی متن را برای پردازش به شبکه داده می شود به کمک روش LDA و TF-IDF برای هر متن پر تکرار ترین کلمات را برای هر متن به صورت جداگانه ایجاد می کند؛ سپس برای هر متن جدیدی که وارد شبکه می شود الگوریتم طراحی شده احتمال شباهت متن را به هر یک از موضوعات بررسی می کند .



```

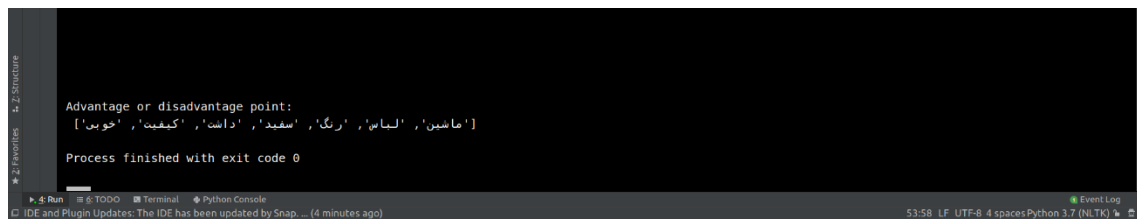
>>>import pandas as pd
>>>from sklearn.feature_extraction.text import CountVectorizer
>>>from sklearn.decomposition import LatentDirichletAllocation
>>>import random
>>>from hazm import *
>>>df=pd.read_csv('title-disadvantage.csv')
>>>cv=CountVectorizer()
>>>dtm=cv.fit_transform(df['disadvantage'])
>>>LDA=LatentDirichletAllocation(n_components=1,random_state=100)
>>>LDA.fit(dtm)
>>>single_topic=LDA.components_[0]
>>>single_topic.argsort()
>>>top_ten_words=single_topic.argsort()[-10:]
>>>topic_dictionary_dis={}
>>>for i,topic in enumerate(LDA.components_):
print(f"THE TOP 15 WORDS FOR TOPIC #{i}")
print([cv.get_feature_names()[index] for index in topic.argsort()[-300:]])
topic_dictionary_dis[i]=[cv.get_feature_names()[index] for index in topic.argsort()[-600:]]
print('\n')
print('\n')
>>>topic_results=LDA.transform(dtm)
>>>df=pd.read_csv('title-advantage.csv')
>>>cv=CountVectorizer()
>>>dtm=cv.fit_transform(df['advantage'])
>>>LDA=LatentDirichletAllocation(n_components=1,random_state=100)
>>>LDA.fit(dtm)
>>>single_topic=LDA.components_[0]
>>>single_topic.argsort()
>>>top_ten_words=single_topic.argsort()[-10:]
>>>topic_dictionary_ad={}
for i,topic in enumerate(LDA.components_):
print(f"THE TOP 15 WORDS FOR TOPIC #{i}")

```

```

print([cv.get_feature_names()[index] for index in topic.argsort()[-300:]])
topic_dictionary_ad[i]=[cv.get_feature_names()[index] for index in
topic.argsort()[-600:]]
print('\n')
print('\n')
>>>topic_results=LDA.transform(dtm)
>>>w='ماشین لباس شویی رنگ سفید جالبی داشت و کیفیت خوبی داشت '
>>>w=w.split(' ')
>>>words=[]
>>>stemmer=Stemmer()
>>>for i in range(len(w)):
words.append(stemmer.stem(w[i]))
>>>dis_ad=topic_dictionary_dis[0]+topic_dictionary_ad[0]
>>>disadvange_advantage=[]
:>>>for i in range(len(dis_ad))
if dis_ad[i] not in disadvange_advantage:
disadvange_advantage.append(dis_ad[i])
>>>point=[]
:>>>for word in w
for i in range(len(dis_ad)):
if word==dis_ad[i]:
if word not in point:
point.append(word)
else:
continue
>>>print("\n\n\n\n\n\nAdvantage or disadvantage point:\n",point)

```



```

Advantage or disadvantage point:
['ماشین', 'لباس', 'رنگ', 'سفید', 'داشت', 'کیفیت', 'خوب']

Process finished with exit code 0

```

شکل ۱۷ خروجی مدل سازی مبحث

همانطور که قبلاً ذکر شد در این نوع تحلیل بررسی دقت خروجی وجود ندارد و بدون معنا می باشد. خروجی ای که در اینجا مشاهده می شود با توجه به ورودی متن زیر:

- "ماشین لباس شویی رنگ سفید جالبی داشت و کیفیت خوبی داشت"

در خروجی دسته کلمات

ماشین ، لباس،رنگ،سفید، کیفیت، خوب

مشاهده می شود که رنگ و سفید و کیفیت ویژگی های مثبت این کالای خریداری شده می باشد .

#### ۴-۶- نتیجه و جمع بندی نهایی:

هدف کلی مقاله همانطور که در ابتدا بیان شد پردازش زبان طبیعی در فارسی بود که سعی شد این تحلیل بر روی داده های آزاد شده دیجیکالا که شامل ۱۵ هزار نظر می باشد انجام شود در حقیقت سعی شد به کمک دو روش مدلی طراحی کنیم که قادر به درک و تحلیل نظر جدید باشد. توقعی که پس از آموزش این مدل داشتیم این بود که مدل ما نظر جدیدی را دریافت کند و پس از تحلیل پیش بینی کند که نظر این کاربر در مورد خرید کالا مثبت است یا منفی به معنای دیگر آیا کاربر از خرید این کالا راضی بوده است یا خیر و خرید این کالا را به دیگر کاربران پیشنهاد می کنند یا پیشنهاد نمی کند و اگر کاربر خرید این کالا را پیشنهاد می کند نکته مثبت این کالا از نگاه خریدار چیست و اگر کاربر خرید کالا را پیشنهاد نمی کند نکته ی منفی کالا از نگاه کاربر چیست. شاید این سوال پیش آید که کاربرد این موضوع چیست؟ تحلیل نظرات کاربران می تواند در راستای افزایش فروش کالا بسیار موثر واقع شود؛ برای مثال اگر کاربر خرید کالا را به علت قیمت بالای آن پیشنهاد نمی کند فروش کالا را با شرایط تخفیف برای کاربران عرضه کنیم و حتی میتوانیم با طراحی شبکه ای تبلیغاتی کاربرانی را که از خرید این کالا ناراضی بودن با پیشنهاد هایی در راستای بهبود نکات منفی کالا آن ها را به خرید کالا ترغیب کرد و اینگونه فروش را افزایش دهیم . برای مثال اگر کاربر از قیمت ناراضی بوده با پیشنهاد تخفیف های ویژه کاربر را به خرید مجدد کالا را ترغیب کنیم در حقیقت با بهبود هر نکته ی منفی از کالا و آگاه کردن خریدار از بر طرف شدن آن نکته آن ها را مجدد به سمت خرید کالا سوق دهیم .

<https://www.edureka.co/blog/types-of-artificial-intelligence>

<https://www.edureka.co/blog/what-is-machine-learning>

<https://www.edureka.co/blog/what-is-deep-learning>

[www.udemy.com \(Udemy.NLP.Natural.Language.Processing.with.Python \)](https://www.udemy.com/course/natural-language-processing-with-python/)



دانشگاه اصفهان

دانشکده فنی - مهندسی

گروه مهندسی برق

## فرم پیشنهاد موضوع پایان نامه کارشناسی

### <موضوع اولیه پیشنهادی>

پژوهشگر: <نام و نام خانوادگی دانشجو> استاد راهنما: دکتر <نام و نام خانوادگی استاد>

نوع پروژه: ☐ عملی ☐ شبیه سازی ☐ مروری ☐ سایر موارد ....

<متن پیشنهاد موضوع> این قسمت باید محدود به ۳۰۰ کلمه باشد. در ابتدا مختصری از تاریخچه و اهمیت موضوع مورد نظر باید ذکر شود. در ادامه، متن پیشنهاد موضوع می تواند به ترتیب عنوان شده تکمیل گردد: روش تحقیق (کارهایی که قرار است صورت بگیرد تا نتیجه مطلوب حاصل گردد)، نتایجی که انتظار می رود حاصل گردد، سؤال هایی که انتظار می رود به آن ها پاسخ داده شود، کاربرد نتایجی و یا محصولی که قرار است حاصل گردد و خلاقیت و نوآوری صورت گرفته در تحقیق.

تجهیزات، امکانات و قطعات پیش بینی شده مورد نیاز:

هزینه مورد نیاز جهت تکمیل پروژه:

فهرست منابع و برخی از مراجع:

نام و نام خانوادگی دانشجو: امضاء: تاریخ:

نام و نام خانوادگی استاد: امضاء: تاریخ:

پیشنهاد پروژه فوق □ با تأخیر ... ماه □ بدون تأخیر در جلسه مورخ ..... شورای گروه  
مهندسی برق مطرح و به تصویب اعضاء رسید.

۱- مدیر گروه مهندسی برق: امضاء: تاریخ:

۲- استاد پروژه: امضاء: تاریخ:

۳-

۴-



دانشگاه اصفهان

دانشکده فنی- مهندسی

گروه مهندسی برق

اطلاعیه برگزاری سمینار پایان نامه کارشناسی

<موضوع پروژه>

پژوهشگر: <نام و نام خانوادگی دانشجو> استاد راهنما: دکتر <نام و نام خانوادگی استاد>

چکیده:

<چکیده پایان نامه>





دانشگاه اصفهان

دانشکده فنی - مهندسی

گروه مهندسی برق

## فرم ارزیابی پایان نامه کارشناسی

بدین وسیله پروژه کارشناسی گروه مهندسی برق با عنوان ... به تلاش پژوهشگر(ان) ... ، راهنمایی ... و داوری ... در تاریخ ... ارائه و مورد ارزیابی قرار گرفته و نتیجه ارزیابی به شرح زیر است:

## نتیجه ارزیابی استاد راهنما:

- |          |   |
|----------|---|
| (۶)....  | کیفیت تحقیق انجام شده توسط دانشجو در زمینه مورد نظر |
| (۴)....  | میزان خلاقیت دانشجو در زمینه تعریف شده              |
| (۴)....  | نحوه تنظیم گزارش                                    |
| (۴)....  | نحوه ارائه پایان نامه                               |
| (۲)....  | نحوه پاسخ گویی به سؤالات                            |
| (۲۰).... | نمره استاد راهنما                                   |

امضاء استاد راهنما

## نتیجه ارزیابی استاد داور:

- |          |   |
|----------|---|
| ....     | آیا متن پایان نامه مطابقت با پروپوزال دارد؟         |
| (۶)....  | کیفیت تحقیق انجام شده توسط دانشجو در زمینه مورد نظر |
| (۴)....  | میزان خلاقیت دانشجو در زمینه تعریف شده              |
| (۴)....  | نحوه تنظیم گزارش                                    |
| (۴)....  | نحوه ارائه پایان نامه                               |
| (۲)....  | نحوه پاسخ گویی به سؤالات                            |
| (۲۰).... | نمره داور   |

.... (۱ نمره به ازای هر مقاله یا محصول) آیا در نتیجه پروژه محصول عملی و یا مقاله ارسال شده به دست آمده؟

.... (۰/۵ - نمره به ازای هر ماه تأخیر) آیا فرم پیشنهاد پروژه بدون تأخیر تصویب شده است؟

.... (۰/۵ - نمره در صورت تأخیر) آیا نسخه اولیه پایان نامه و اطلاعیه در زمان مربوطه آماده شده است؟

امضاء استاد داور

امضاء مدیر گروه

نتیجه ارزیابی نهایی (۱۵ نمره استاد راهنما، ۵ نمره استاد داور): ...

\* این برگه به همراه فرم تصویب شده پیشنهاد پروژه به مدیر گروه تحویل داده شود.