

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



پردیس بین المللی خوارزمی  
رساله دکتری مهندسی هوش مصنوعی

## ارائه یک مدل آماری جهت ارزیابی سیستم‌های پرسش و پاسخ تعاملی

نگارنده:

محمد مهدی حسینی

استاد راهنما:

دکتر مرتضی زاهدی

استاد مشاور:

دکتر حمید حسن پور

تیر ۱۳۹۷



شماره: ۰۲/۲۷۲۴۹ الف

تاریخ: ۱۱/۱۲/۹۷

ویرایش:

باسمه تعالی



مدیریت تحصیلات تکمیلی

فرم شماره ۱۲: صورت جلسه نهایی دفاع از رساله دکتری (Ph.D)  
(ویژه دانشجویان ورودی های ۹۴ و ما قبل)

بدینوسیله گواهی می شود آقای/خانم محمدمهدی حسینی دانشجوی دکتری رشته کامپیوتر- هوش مصنوعی به شماره دانشجویی ۹۲۴۶۳۴۵ ورودی بهمن ماه سال ۱۳۹۲ در تاریخ ۹۷/۰۴/۱۰ از رساله نظری / عملی خود با عنوان : ارائه یک مدل آماری جهت ارزیابی سیستم های پرسش و پاسخ تعاملی دفاع و با اخذ نمره ۱۸/۷۸ به درجه بسیار خوب رسیده است. این نمره نائل گردید.

<input checked="" type="checkbox"/> الف) درجه عالی: نمره ۱۹-۲۰	<input type="checkbox"/> ب) درجه بسیار خوب: نمره ۱۸/۹۹ - ۱۷
<input type="checkbox"/> ج) درجه خوب: نمره ۱۶/۹۹ - ۱۵	<input type="checkbox"/> د) غیر قابل قبول و نیاز به دفاع مجدد دارد
<input type="checkbox"/> ه) رساله نیاز به اصلاحات دارد	

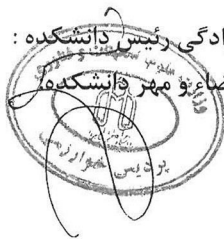
ردیف	هئیت داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
	دکتر مرتضی زاهدی	استاد/ اساتید راهنما		
	دکتر حمید حسن پور	مشاور / مشاورین		
	دکتر مروی	استاد مدعو داخلی / خارجی		
	دکتر قاسم زاده	استاد مدعو داخلی / خارجی		
	دکتر مشایخی	استاد مدعو داخلی / خارجی		
	دکتر عباس نژاد	سرپرست ( نماینده ) تحصیلات تکمیلی دانشکده		

مدیر محترم تحصیلات تکمیلی دانشگاه:

ضمن تأیید مراتب فوق مقرر فرمائید اقدامات لازم در خصوص انجام مراحل دانش آموختگی آقای/خانم محمد مهدی حسینی بعمل آید.

نام و نام خانوادگی رئیس دانشکده:

تاریخ و امضاء و مهر دانشکده:



تقدیم به خانوادهام کہ ہمیشہ یاور من بوده‌اند

## تقدیر و تشکر

تشکر و قدردانی از دکتر مرتضی زاهدی به عنوان استاد راهنما که راهنمایی‌ها و تشویق‌های ایشان سبب گردید که بتوانم این رساله را به پایان برسانم و کمال تشکر از دکتر حمید حسن‌پور که با راهنمایی‌های دلسوزانه‌شان من را در تکمیل این رساله یاری نمودند، دارم.

از دوستان و همکاران عزیز آزمایشگاه تکنولوژی زبان‌های طبیعی دانشگاه صنعتی شاهرود به ویژه سرکار خانم مرضیه رحیمی، آقایان مهدی یعقوبی، محمد عبداللهی و ایمان فیروزیان که در جلسات و بحث‌های انجام شده، راهگشای مسائل مطرح شده در این رساله بودند، کمال تشکر را دارم. همچنین از کلیه دانشجویان دانشکده کامپیوتر دانشگاه صنعتی و آزاد که در ساختن پایگاه داده با اینجانب همکاری داشتند، سپاسگزار هستم.

## تعهد نامه

اینجانب محمد مهدی حسینی دانشجوی دوره دکتری رشته مهندسی کامپیوتر-هوش مصنوعی دانشکده کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده پایان نامه ارائه یک مدل آماری جهت ارزیابی سیستم‌های پرسش و پاسخ تعاملی تحت راهنمایی آقای دکتر مرتضی زاهدی متعهد می شوم .

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

### تاریخ

### امضای دانشجو

#### مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است ) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

## چکیده:

سیستم پرسش و پاسخ، یک سیستم خودکار جهت پاسخ به سؤالاتی است که توسط انسان در زبان طبیعی مطرح می‌شود. در این سیستم‌ها در صورتی که پاسخ مورد انتظار کاربر نبوده و یا کاربر، نیاز به دریافت اطلاعات بیشتر داشته باشد، امکان تبادل اطلاعات، بین سیستم و کاربر وجود ندارد. بنابراین برای رفع این معضل، سیستم‌های پرسش و پاسخ تعاملی بوجود آمدند. ارزیابی نقش مهمی در ارتقای سیستم‌های پرسش و پاسخ تعاملی ایفا می‌نماید. تحقیقات صورت گرفته در این زمینه نشان داده، تاکنون روش خاصی که به ارزیابی کلی این سیستم‌ها بپردازد، وجود نداشته و تنها امکان بهره‌گیری از روش‌های ارزیابی در سیستم‌های پرسش و پاسخ، سیستم‌های گفت‌وگو محور و استفاده از ارزیاب انسانی میسر می‌باشد. لذا ارائه یک راهکار مناسب جهت خودکار کردن روند ارزیابی سیستم‌های پرسش و پاسخ تعاملی، می‌تواند کمک شایانی به ارتقای این سیستم‌ها نماید.

ارائه یک مدل به جای ارزیاب انسانی می‌تواند یک کار چالش برانگیز برای ارزیابی سیستم‌های تعاملی باشد بطوریکه خروجی مدل بتواند امتیاز داده شده توسط ارزیاب را پیش‌بینی نماید. بنابراین در این رساله تلاش شده است تا با تعیین و انتخاب پارامترهای مناسب در ارزیابی، یک مدل آماری برای ارزیابی سیستم‌های پرسش و پاسخ تعاملی ارائه گردیده تا در روند ارزیابی مورد استفاده قرار گیرد. هدف از ارائه مدل آماری، عدم وابستگی مدل به زبان سیستم تعاملی می‌باشد. در این راستا به منظور رسیدن به مناسب‌ترین مدل، چندین ویژگی آماری ایجاد و با استفاده از رگرسیون و برنامه‌نویسی بیان ژن به استخراج مدل پرداختیم. ابتدا، یک پایگاه داده از مکالمات صورت گرفته با چهار سیستم پرسش و پاسخ تعاملی ایجاد گردید. سپس استخراج ویژگی بر روی هر مکالمه صورت پذیرفت و در نهایت با استفاده از رگرسیون به استخراج مدل پرداختیم. با توجه به تعداد بالای ویژگی‌های پیشنهادی و برای جلوگیری از برازش خطا، بهترین ویژگی‌ها با استفاده از روش حذف ویژگی به روش بازگشتی انتخاب گردید تا مدل پیشنهادی بر اساس ویژگی‌های باقیمانده شکل گیرد. بهترین مدل ارائه شده با توجه به

معیار ارزیابی جذر میانگین مربعات خطا، توسط رگرسیون لسو و سری توانی حاصل گردید. در ادامه برای رسیدن به مدل مناسب‌تر، برنامه‌نویسی بیان ژن مورد استفاده قرار گرفت. در اولین مرحله، یک معادله رگرسیونی جهت پیش‌بینی امتیاز ارزیاب‌ها براساس کل مکالمات صورت پذیرفت. در ادامه مکالمات به سه کلاس با امتیاز خوب، متوسط و ضعیف دسته‌بندی شدند و برای هر کلاس یک معادله رگرسیونی حاصل گردید. بنابراین برای یک مکالمه جدید، بعد از مرحله استخراج ویژگی و تخصیص آن به یک کلاس، می‌توان امتیاز را بر اساس مدل رگرسیونی، محاسبه نمود. با توجه به ارزیابی‌های صورت گرفته بر روی مدل، میزان میانگین خطای خروجی مدل نسبت به خروجی واقعی برابر با ۰/۰۹ حاصل گردید که نشان از ارائه مدل مناسب به منظور ارزیابی این سیستم‌ها می‌باشد. روش ارائه شده در این پژوهش نشان داد که اگر یک مجموعه استاندارد از مکالمات صورت گرفته بین کاربران و سیستم موجود باشد، می‌توان در صورت عدم وجود ارزیاب انسانی، امتیاز حاصل از ارزیابی را برای سیستم پرسش و پاسخ تعاملی پیش‌بینی نمود.

**کلمات کلیدی:** سیستم پرسش و پاسخ تعاملی، ارزیابی سیستم‌ها، سیستم پرسش و پاسخ،

رگرسیون، برنامه‌نویسی بیان ژن

## مقالات مستخرج از رساله

مقالات ارائه شده در نشریات علمی - پژوهشی			
ردیف	عنوان مقاله	وضعیت	تاریخ
۱	محمد مهدی حسینی، مرتضی زاهدی "ارائه یک مدل احتمالاتی جهت تعیین انسجام متن در سیستم‌های پرسش و پاسخ تعاملی" - مجله پردازش و مدیریت اطلاعات	چاپ شده	دی ۹۶
۲	محمد مهدی حسینی، مرتضی زاهدی، حمید حسن پور، "مجموعه ای از ویژگی‌های آماری جدید برای ارزیابی سیستم‌های پرسش و پاسخ تعاملی" - مجله مهندسی برق تبریز	چاپ شده	اردیبهشت ۹۷
مقالات ارائه شده در کنفرانس‌ها			
۱	محمد مهدی حسینی، مرتضی زاهدی "استخراج پاسخ مناسب به تقاضا براساس شباهت معنایی در سیستم‌های پاسخگویی اتوماتیک" اولین کنفرانس پردازش سیگنال و سیستم‌های هوشمند	چاپ شده	آذر ۹۴
۲	محمد مهدی حسینی، مرتضی زاهدی، "بهبود پاسخ ارائه شده در سیستم‌های پرسش و پاسخ تعاملی به کمک شبکه عصبی"، هشتمین کنفرانس بین المللی فناوری اطلاعات و دانش	چاپ شده	شهریور ۹۵
۳	محمد مهدی حسینی، مرتضی زاهدی "ارائه مجموعه‌ای از ویژگی‌های آماری جهت تعیین انسجام متن در سیستم‌های پرسش و پاسخ تعاملی"، سومین کنفرانس پردازش سیگنال و سیستم‌های هوشمند	چاپ شده	آذر ۹۶

## علائم اختصاری

BOW	Bag of Words
CQA	Community Question Answering
CIEF	Conference and Lab Evaluation Forum
FQS	Follow-up Questions
GEP	Gene Expression Programming
GP	Genetic programming
IQA	Interactive Question Answering
LSTM	Long Short Term Memory
MAPE	Mean Absolute Percentage Error
NLP	Natural language processing
QA	Question Answering
RMSE	Root Mean Square Error
$R^2$	R-squared
TREC	Text Retrieval Evaluation Conference



فصل اول - مقدمه.....	۱
۱-۱ سیستم‌های پرسش و پاسخ.....	۲
۲-۱ سیستم‌های پرسش و پاسخ تعاملی.....	۵
۳-۱ طرح مسئله.....	۷
۴-۱ چالش‌های موجود در فرآیند ارزیابی.....	۱۰
۵-۱ فرضیه سیستم ارزیابی.....	۱۲
فصل دوم- سابقه تحقیق.....	۱۵
۱-۲ روند پیدایش سیستم‌های پرسش و پاسخ.....	۱۶
۲-۲ مروری بر کارهای صورت پذیرفته در سیستم‌های پرسش و پاسخ تعاملی.....	۱۹
۳-۲ روش‌های ارزیابی سیستم‌های پرسش و پاسخ تعاملی.....	۲۴
۴-۲ معیارهای ارزیابی موجود.....	۳۷
فصل سوم- روش پیشنهادی.....	۴۵
۱-۳ روش و طرح کلی تحقیق.....	۴۶
۲-۳ پیش‌پردازش.....	۴۸
۳-۳ استخراج ویژگی.....	۴۹
۴-۳ تعیین انسجام متن در سیستم‌های IQA با ارائه یک مدل احتمالاتی.....	۶۱
۵-۳ مدل‌سازی داده‌ها با استفاده از آنالیز رگرسیونی.....	۶۳
۶-۳ برنامه‌نویسی بیان ژن جهت تعیین مدل مناسب.....	۶۹
۷-۳ یادگیری عمیق.....	۷۱
۸-۳ معیار ارزیابی.....	۷۲
فصل چهارم- تجزیه و تحلیل یافته‌های تحقیق.....	۷۵
۱-۴ مجموعه داده.....	۷۶
۲-۴ آزمایش شماره ۱: نتایج حاصل از تعیین انسجام متن در سیستم‌های IQA.....	۸۰
۳-۴ آزمایش شماره ۲: یافتن بهترین مدل بر اساس ویژگی‌های پیشنهادی به کمک رگرسیون.....	۸۴
۴-۴ آزمایش شماره ۳: ارزیابی ویژگی‌های پیشنهادی.....	۱۰۰
۵-۴ آزمایش شماره ۴: استخراج مدل رگرسیونی به کمک برنامه‌نویسی بیان ژن.....	۱۰۲
۶-۴ آزمایش شماره ۵: استفاده از شبکه LSTM در یادگیری عمیق جهت پیش‌بینی امتیاز کاربر.....	۱۱۰
فصل پنجم- نتیجه‌گیری و کارهای آینده.....	۱۱۵
۱-۵ جمع‌بندی روش پیشنهادی در رساله.....	۱۱۶
۲-۵ کارهای آینده.....	۱۲۰
منابع.....	۱۲۳

۱۲۹	پیوست الف-رگزیون.....
۱۴۱	پیوست ب-برنامه نویسی بیان ژن.....
۱۴۹	پیوست ج-یادگیری عمیق.....



شکل ۱-۱	فرآیند کلی یک سیستم QA.....	۳
شکل ۲-۱	نمای مربوط به فرآیند مدل‌سازی در رساله.....	۹
شکل ۱-۳	فرآیند شکل‌گیری روش ارزیابی پیشنهادی.....	۴۷
شکل ۲-۳	تشریح مراحل روش پیشنهادی برای تعیین انسجام متن در سیستم‌های IQA.....	۶۳
شکل ۳-۳	بررسی روند انتخاب نوع مدل رگرسیونی با توجه به متغیرهای ورودی.....	۶۶
شکل ۴-۳	نمایی قسمتی از ماتریس ویژگی به همراه نمودار پراکندگی برخی از ویژگی‌ها جهت نمایش عدم امکان حدس تابع.....	۶۷
شکل ۱-۴	نمایی از فرآیند استفاده از صفحات مختلف سامانه تحت وب طراحی شده برای ارزیابی.....	۷۷
شکل ۲-۴	تابع توزیع تجمعی ۳۲ نقطه‌ای ویژگی اول برای یک مکالمه و سه تابع توزیع تجمعی تقریب زده شده.....	۸۱
شکل ۳-۴	نمودار میله‌ای و پراکندگی برخی از ویژگی‌ها.....	۸۵
شکل ۴-۴	ضرایب معادله رگرسیون خطی چندگانه.....	۸۸
شکل ۵-۴	نتایج حاصل از رگرسیون خطی چندگانه.....	۸۹
شکل ۶-۴	نتایج حاصل از انتخاب مدل با استفاده از رگرسیون چندجمله‌ای.....	۹۱
شکل ۷-۴	فرآیند استفاده از رگرسیون لسو.....	۹۴
شکل ۸-۴	روند پیچیدگی مدل در برابر معیارهای ارزیابی برای مجموعه آموزش و تست با استفاده از رگرسیون لسو.....	۹۵
شکل ۹-۴	ضرایب معادله رگرسیون غیرخطی توانی.....	۹۶
شکل ۱۰-۴	ضرایب معادله رگرسیون غیرخطی سری توانی.....	۹۷
شکل ۱۱-۴	تعیین ضریب تاثیر هر ویژگی با استفاده از رگرسیون جنگل تصادفی.....	۹۸
شکل ۱۲-۴ (الف)	معادله خط به دست آمده برای مکالمات مربوط به کلاس خوب (ب) معادله خط به دست آمده برای مکالمات مربوط به کلاس متوسط (ج) معادله خط به دست آمده برای مکالمات مربوط به کلاس ضعیف.....	۱۰۶
شکل ۱۳-۴	تصدیق مدل به دست آمده برای کلاس خوب با استفاده از داده‌های آزمون.....	۱۰۹
شکل ۱۴-۴	مقادیر برخی از ویژگی‌ها به دست آمده جهت پیش‌بینی امتیاز تخصیص شده با استفاده از LSTM.....	۱۱۳
شکل ۱۵-۴	نمودار مربوط به مقادیر از دست رفته مجموعه آموزش و آزمون در طول فرآیند آموزش.....	۱۱۴
شکل ب-۱	طرح کلی گام‌های مقدماتی برنامه نویسی بیان ژن [۵۶].....	۱۴۳
شکل ب-۲	فلوچارت برنامه‌ریزی بیان ژن [۵۷].....	۱۴۶
شکل ب-۳	کدگذاری به فرم رشته خطی و کد برداری به صورت ET کروموزوم با دو ژن در GEP [۵۵].....	۱۴۷
شکل ج-۱	ساختار یک شبکه عصبی بازگشتی [۵۸].....	۱۵۱
شکل ج-۲	ساختار شبکه LSTM در برابر شبکه RNN [۵۸].....	۱۵۲
شکل ج-۳	بلوک دیاگرام مربوط به LSTM [۵۸].....	۱۵۳
شکل ج-۴	ساختار یک گیت [۵۸].....	۱۵۴

جدول ۱-۱	مقایسه بین رویکردهای NLP محور با قانون محور [۱]	۴
جدول ۲-۱	مقایسه کلی بین دسته‌های مختلف موجود برای سیستم‌های پرسش و پاسخ [۲]	۶
جدول ۳-۱	روند پیشرفت سیستم‌های QA	۲۰
جدول ۱-۳	لیست ویژگی‌های در نظر گرفته شده جهت ارزیابی سیستم‌های پرسش و پاسخ	۵۱
جدول ۱-۴	یک نمونه تولید شده از یک مکالمه در سیستم IQA	۷۸
جدول ۲-۴	یک نمونه از مکالمه تولید شده در سیستم BOBDIALOGUE	۷۸
جدول ۳-۴	یک نمونه از مکالمه مناسب و نامناسب	۷۹
جدول ۴-۴	یک نمونه از داده آموزشی استفاده شده در آموزش سیستم تعاملی پایه	۸۰
جدول ۵-۴	نتایج تطبیق سه تابع چگالی احتمال با کل ویژگی‌ها در سه گروه مکالمه انتخاب شده	۸۲
جدول ۶-۴	بررسی تطبیق یک کلاس از مکالمه با خودش در سه تابع چگالی احتمال مورد بررسی	۸۲
جدول ۷-۴	مقایسه بین تابع چگالی احتمال کلاس‌های مکالمه مختلف و آزمون‌های آماری	۸۳
جدول ۸-۴	نمایش برخی از ضرایب بدست آمده برای دستیابی به معادله خط رگرسیونی	۹۲
جدول ۹-۴	مقایسه نتایج رگرسیون ریج با چهار روش دیگر	۹۳
جدول ۱۰-۴	ضرایب همبستگی پیرسون برای پایگاه داده ایجاد شده	۱۰۱
جدول ۱۱-۴	مقادیر اولیه برای برنامه‌نویسی بیان ژن	۱۰۴
جدول ۱۲-۴	شرایط شبیه‌سازی به کمک برنامه‌نویسی بیان ژن	۱۰۴
جدول ۱۳-۴	مقادیر به دست آمده برای پیش‌بینی نظرات کاربران به تعداد ۵ نمونه برای هر کلاس	۱۰۹



## فصل اول - مقدمه

سیستم پرسش و پاسخ

سیستم پرسش و پاسخ تعاملی

طرح مسئله

در دنیای مملو از اطلاعات، دستیابی به پاسخ‌های صحیح و دقیق در کوتاه‌ترین زمان ممکن، به یکی از تکنولوژی‌های مورد نیاز کاربران دنیای مجازی مبدل شده است. در این راستا، کاربران اینترنتی تلاش دارند تا با بهره‌گیری از ابزارهای موجود در دنیای وب مانند موتورهای جستجو یا گوگل، نیازهای اطلاعاتی خود را در حداقل زمان و با حداکثر سرعت ممکن برطرف نمایند. از این رو، سیستم‌های بازیابی اطلاعات متنوعی برای رسیدن به پاسخ‌سؤالات کاربران، در قالب متن، صوت، تصویر، ویدئو یا ترکیبی از آن‌ها طراحی و ارائه شده‌اند [۱].

سیستم‌های پرسش و پاسخ متنی (QA)<sup>۱</sup>، یک از روش‌های بازیابی اطلاعات است که در ازای پرسش مطرح شده، پاسخ مستقیمی را از میان اسناد مرتبط استخراج و در قالب لیستی از جملات و یا قطعات متنی، به کاربر نمایش می‌دهند [۲]. وجود ابهام در پاسخ ارائه شده و یا غیرقابل قبول بودن پاسخ دریافت شده توسط کاربران در بعضی از موارد، سبب گردید تا سیستم‌های پرسش و پاسخ تعاملی (IQA)<sup>۲</sup> به وجود آیند. از مسائل چالش برانگیز در حوزه سیستم‌های IQA، ارزیابی این سیستم‌ها است. در این فصل به تشریح سیستم‌های QA، IQA و طرح مسئله ارزیابی آن‌ها پرداخته‌ایم.

## ۱-۱ سیستم‌های پرسش و پاسخ

سیستم QA به عنوان سیستمی با قابلیت بالا شناخته می‌شود که کاربران را قادر می‌سازد تا به اسناد مرتبط با سوال مطرح شده، دسترسی داشته و پاسخی مناسب و مختصر برای آن دریافت نمایند. یکی از وظایف مهم سیستم‌های QA، درک صحیح سؤالات مطرح شده به زبان طبیعی (NLP)<sup>۳</sup> و فهم معنی دقیق کلمات سؤال، جهت ارائه پاسخ مناسب است. چالش‌های متعددی در این حوزه وجود دارد که از جمله آن‌ها می‌توان به بهبود درک ماشین از سؤالات، طبقه‌بندی، فرمول‌بندی صحیح پرس‌وجوها، ابهام در تجزیه، تشخیص تقارن معنایی، تشخیص جواب مناسب برای کاربر، مکانیسم اعتبارسنجی مناسب و تشخیص روابط ظاهری در سؤالات پیچیده، اشاره نمود [۱، ۳]. ساختار کلی یک

<sup>۱</sup> Question Answering

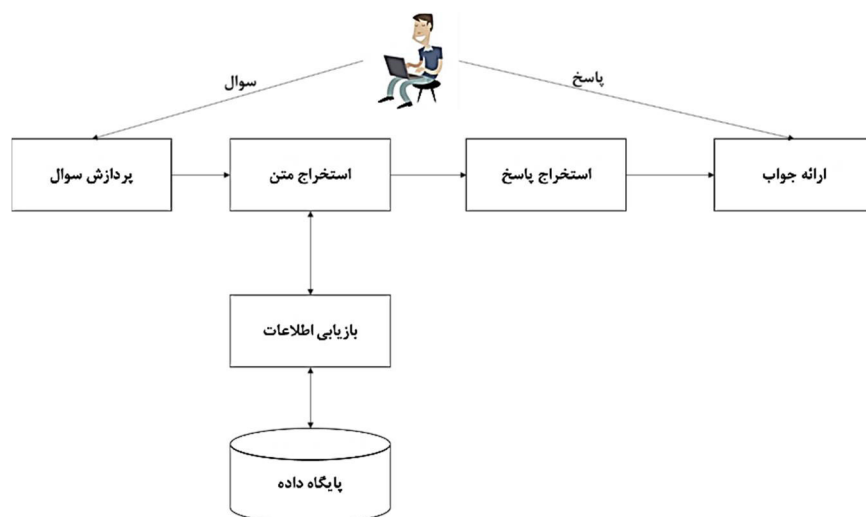
<sup>۲</sup> Interactive Question Answering

<sup>۳</sup> Natural Language Processing



سیستم QA در شکل ۱-۱ نشان داده شده است. یک سیستم QA از سه بخش زیر تشکیل شده است.

- پردازش سؤال: تجزیه نحوی، کلاس‌بندی سؤال و فرمول‌بندی مجدد پرسش وجود.
  - تحلیل متن: استخراج متون مناسب و تشخیص پاسخ بر اساس بازیابی اطلاعات از پایگاه داده موجود.
  - تحلیل پاسخ: استخراج جواب‌های مناسب و رتبه‌بندی بهترین پاسخ و ارائه آن به کاربر.
- در جدیدترین پژوهش‌ها، برای ایجاد سیستم‌های یکپارچه در فرآیند پرسش و پاسخ، از برخی تکنیک‌های هوش مصنوعی مانند پردازش زبان طبیعی، تطابق الگو، بازیابی اطلاعات، استخراج اطلاعات و یا ترکیب آن‌ها بهره گرفته شده است. بنابراین بر اساس رویکرد در نظر گرفته شده، سیستم‌های QA که تاکنون توسعه یافته‌اند را می‌توان به سه گروه رویکرد زبان‌شناختی، آماری و تطابق الگو طبقه‌بندی نمود [۱].



شکل ۱-۱ فرآیند کلی یک سیستم QA

## ۱-۱-۱ رویکرد زبان‌شناختی

محققین این حوزه با استفاده از تکنیک‌های متنوع، سعی در طراحی سیستم‌های پرسش و پاسخ کارآمد و دقیق داشته‌اند. بنابراین در این سیستم‌ها با ترکیب تکنیک‌های پردازش زبان، ساخت پایگاه‌های دانش، تجزیه و تحلیل معنایی و گرامری، سعی شد تا پرسش کاربر به یک پرس‌وجوی ساخت‌یافته تبدیل شود. در این روش‌ها، تهیه پایگاه دانش، زمان بر و کاربرد سیستم به دامنه‌های موضوعی خاص محدود می‌باشد. جدول ۱-۱ تفاوت بین سیستم‌های QA با محوریت پردازش زبان طبیعی و قانون را نمایش می‌دهد.

## ۱-۱-۲ رویکرد آماری

سیستم‌های موجود در این دسته مستقل از پایگاه داده می‌باشند و می‌توانند پرس‌وجوها را به فرم زبان طبیعی فرمول‌بندی نمایند. رشد سریع داده‌های آنلاین و قابل دسترس بودن متن و داده‌های وب، موجب افزایش اهمیت این سیستم‌ها گردیده است [۱]. این رویکرد، تکنیک‌هایی را پیشنهاد می‌کنند که قابلیت مدیریت حجم زیاد و متنوعی از داده را دارند.

جدول ۱-۱ مقایسه بین رویکردهای NLP محور با قانون محور [۱]

قانون	NLP	
متون ساختاریافته یا نیمه ساختارمند که به‌طور کلی برای سیستم‌های ادراک محور گسترش یافته‌اند.	پایگاه دانش ساختارمند	سازمان‌دهی منابع
قوانین اکتشافی تفسیری برای تشخیص کلاس‌ها	عدم نیاز	تفسیر کلاس پرسش
تکنیک‌های NLP سطحی	تکنیک‌های NLP عمیق	تحلیل زبان‌شناختی
محدود به مفاهیم و یا متون از قبل ذخیره شده	اکثراً کوچک بوده اما با استفاده از وب به عنوان منبع علمی گسترش یافته	حوزه و گستره دانش
نیاز به یادگیری از طریق داده‌های آموزشی	مورد نیاز نیست	فراگیری اطلاعات

آنچه که می‌بایست در این سیستم‌ها مورد توجه قرار گیرد، ضرورت آموزش آن‌هاست که باید با استفاده از حجم بزرگی از داده‌های آموزشی صورت پذیرد. این روش‌ها تنها در صورت آموزش صحیح، توانایی پاسخ‌گویی دقیق را خواهند داشت. این سیستم‌ها مستقل از ساختار و شکل زبانی هستند و برنامه یا روش آموزش آن‌ها می‌تواند به راحتی برای یک حوزه جدید تعمیم داده شود. یکی از نقاط ضعف اصلی این دسته از سیستم‌ها، رفتار مستقل با هر واژه است که سبب می‌شود تا در تشخیص خصوصیات زبان‌شناختی برای ترکیبی از کلمات و عبارات، ناموفق عمل نمایند.

### ۱-۱-۳ رویکرد تطابق الگو

در این رویکرد، قوانین توسط افراد خبره، در حوزه خاصی از دانش تعریف می‌شود. این رویکرد، یکی از تکنیک‌های ساده مبتنی بر دانش و بدون استفاده از روش‌های پیچیده پردازش زبان طبیعی است، که می‌تواند جهت انتخاب مناسب‌ترین پاسخ در سیستم‌های QA استفاده شود. اغلب سیستم‌های QA با نگرش تطابق الگو، از الگوهای ظاهری استفاده می‌کنند. در حالی که، بعضی از آن‌ها همچنان برای ایجاد پاسخ، وابسته به الگوهای قالب‌بندی شده از قبل می‌باشند. جهت ایجاد یک سیستم QA باید به ویژگی‌های مطرح شده برای هر رویکرد توجه نمود و متناسب با آن به طراحی یک سیستم پرداخت. جدول ۱-۲ مقایسه کلی بین این سه رویکرد را نمایش می‌دهد.

### ۱-۲ سیستم‌های پرسش و پاسخ تعاملی

یکی از چالش‌های موجود در سیستم‌های پرسش و پاسخ، فقدان تعامل دوطرفه بین سیستم و کاربر است. افزودن سطح تعامل به سیستم‌های پرسش و پاسخ با دو هدف صورت می‌گیرد. نخست این که اگر پرسش کاربر دارای ابهام باشد، مکالمه‌ای با کاربر در جهت رفع ابهام و درج بهتر پرسش آغاز شود و دوم این که چنانچه پاسخ سیستم، دلخواه کاربر نباشد و یا کاربر نیاز به دریافت اطلاعات بیشتری داشته باشد، کاربر مکالمه‌ای را با سیستم آغاز می‌نماید تا پاسخ دلخواه خود را دریافت کند.

جدول ۱-۲ مقایسه کلی بین دسته‌های مختلف موجود برای سیستم‌های پرسش و پاسخ [۳]

تطابق الگو	آماری	زبان‌شناسی	
حقیقت، تعریفی، دریافت صورت کامل کلمات مخفف، تاریخ تولد	پرسش‌های غیر حقیقت پیچیده به همراه پرسش‌های حقیقت	پرسش‌هایی در مورد حقایق	نوع پرسش
کمتر از دو رویکرد دیگر	کم عمق	عمیق	درک معنایی
به دلیل ماهیت الگوها، به راحتی امکان‌پذیر است.	برای یکپارچه‌سازی داده‌ها از معیارهای شباهت آماری استفاده می‌شود.	نسبتاً دشوار است، زیرا پایگاه‌های دانش تنها برای مدیریت نوع داده‌های از پیش تعیین شده برایشان طراحی شده‌اند.	مدیریت داده‌های ناهمگون
به اعتبار منابع دانش بستگی دارد.	به دلیل استفاده از رویکردهای نظارتی، قابل اعتماد است.	بیشترین قابلیت اعتماد را دارد.	قابلیت اتکا و اعتماد
کم است، چون به ازای هر مفهوم جدید الگوهای جدید باید آموخته شوند.	به محض اینکه به درستی آموزش ببینند برای این کار بسیار مناسب خواهد بود.	چون به ازای هر مفهوم جدید باید قوانین جدید به پایگاه دانش افزوده شوند، لذا نسبتاً پیچیده و دشوار است.	مقیاس‌پذیری
وب‌سایت‌های کوچک و متوسط، وب معنایی	حجم زیاد داده مانند وب	در یک دامنه خاص نیازمندی‌های اطلاعاتی بسیاری وجود داشته باشد	حوزه‌ی کاربرد

سیستم‌های تعاملی به کاربر این امکان را می‌دهند که در ابتدا سؤال خود را پرسیده، سپس با استفاده از بازخورد مربوط به سؤال، سیستم درصدد اصلاح و یا پاسخگویی برآمده، تا کاربر پاسخ مورد نیاز خود را دریافت نماید [۴]. بنابراین می‌توان این‌طور تصور نمود که در IQA یک فرایند تکراری اتفاق می‌افتد بنابراین سیستم‌های IQA از سیستم‌های QA دقیق‌تر هستند [۵]. زمانی که یک سیستم IQA با ساختار مبهمی در پرسش مطرح شده مواجه می‌شود، برای واضح‌تر شدن درخواست، دیالوگی را با کاربر آغاز می‌نماید. به بیان دیگر، این امکان برای سیستم فراهم شده تا در صورت نیاز برای رفع ابهام

در سؤال، بتواند سوالات دیگری را از کاربر بپرسد. سیستم‌های موجود IQA می‌توانند با توجه به شرایط و کاربردهایشان در سه گروه مختلف قرار گیرند [۴].

الف) مدیریت محدودیت: این سیستم‌ها، محدودیت در سؤالات را شناسایی کرده و هنگامی که اصلاح یا تغییر محدودیت‌ها ضروری باشد با کاربر تعامل می‌کنند. معمولاً بخش عمده این پردازش در این‌گونه سیستم‌ها، توسط مدیر دیالوگ (گفتگو) صورت می‌پذیرد.

ب) سیستم پرسش و پاسخ ارتقاء یافته: محققان این حوزه بر شناسایی دقیق IQA و شناسایی دامنه کارکرد آن تکیه دارند. این شناسایی به‌طور کلی جنبه‌های قابل توجهی از تعامل و تأثیرگذاری متقابل را در برمی‌گیرد و بر پایه، آنالیز سؤالات و جواب‌ها بوده و در مدل‌سازی نمونه‌های دیالوگ مورد استفاده قرار می‌گیرند.

ج) سؤالات متوالی: سیستم‌های پاسخگویی به سؤالات متوالی<sup>۱</sup> (FQS) به قلمرو سیستم‌های IQA تعلق دارند. محققان این حوزه بیشتر بر روی متن و روند پردازش در سیستم‌های QA متمرکزند. تکمیل متن در این نوع از سیستم‌ها شامل مراحل شناسایی موارد مورد نیاز برای تکمیل متن، شناسایی استراتژی بازنویسی و تکرار متوالی عبارات و انتخاب ارجح می‌باشد. هر یک از این مراحل چالش‌های خاص خودشان را داشته و باید به دقت مورد توجه قرار گیرند. به عنوان مثال، گاهی اوقات تغییر موضوع، دسته‌بندی صحیح FQS را با مشکل مواجه می‌نماید.

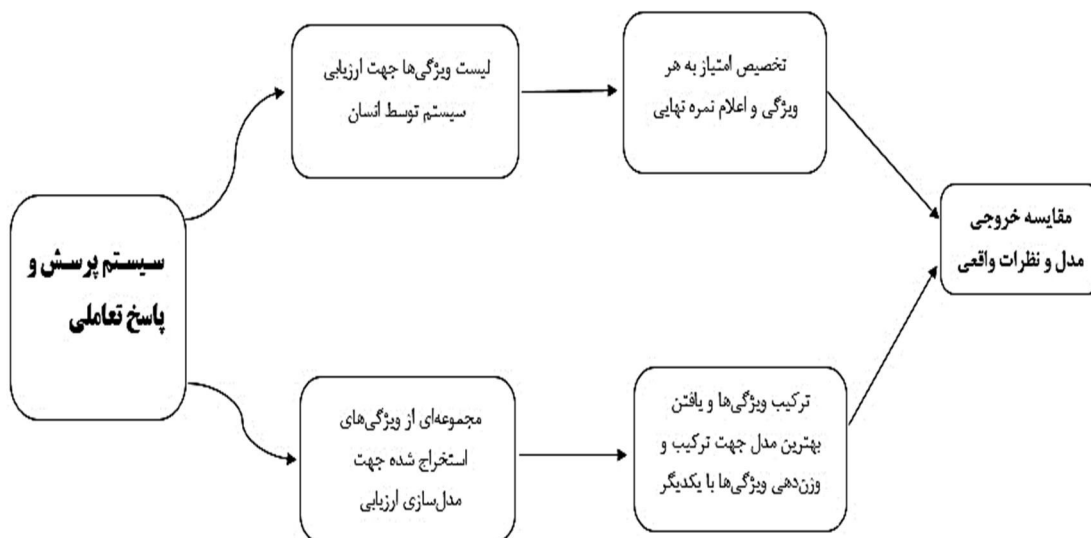
### ۱-۳ طرح مسئله

برای درک درست نسبت به ضرورت انجام این کار اجازه دهید صورت مسئله را با یک مثال ساده بیان کنیم تا بتوان ابعاد ضرورت انجام این کار را روشن نماییم. فرض کنید شما می‌خواهید یک خانه تهیه کنید. قاعدتاً در ابتدا لیستی از ویژگی‌های موثر برای خرید خانه را تهیه و سپس برای خرید اقدام

<sup>۱</sup> Follow-up Questions

می‌نمایید. بر اساس خانه‌هایی که به مشاهده کرده‌اید، به هر کدام از ویژگی‌ها یک امتیاز تخصیص می‌دهید و در نهایت با توجه به همه امتیازات داده شده، به انتخاب یک خانه از بین دیگر خانه‌ها می‌رسید. در حالی که ممکن است بر اساس همان ویژگی‌ها شخص دیگری، خانه دیگری را انتخاب نماید. دلیل این امر در نظر گرفتن امتیاز تخصیص داده شده به هر ویژگی و نحوه ترکیب آن‌ها با یکدیگر است. بنابراین توجه به کشف دانش در این مورد که افراد برای انتخاب یک خانه از چه تابع ارزیابی استفاده کرده‌اند، خود می‌تواند یک چالش باشد. حال اجازه دهید همین مسئله را به ارزیابی سیستم‌های پرسش و پاسخ تعاملی تعمیم دهیم. با توجه به اینکه در حال حاضر در فرآیند ارزیابی سیستم‌های پرسش و پاسخ تعاملی افراد متخصص و خبره نقش دارند. بنابراین حدس اینکه این افراد در پس زمینه ذهن خود از چه تابع ارزیابی برای نمره دهی به یک سیستم استفاده نموده‌اند، از چالش‌های موجود در زمینه ارزیابی این سیستم‌ها است. با به دست آوردن این تابع می‌توان به ارزیابی یک سیستم جدید IQA با کمترین هزینه و در سریع‌ترین زمان ممکن پرداخت. به نکته‌ای که در فرآیند ایجاد مدل باید توجه نمود این است که مدل پیشنهادی برای ارزیابی سیستم‌های IQA باید طوری باشد که خروجی آن کمترین خطا را نسبت به نظرات موجود از سیستم را داشته باشد.

در حقیقت ما در این رساله این هدف را دنبال می‌کنیم که یک مدل ارزیابی پیشنهاد نماییم تا وقتی که موسسه یا فردی خواست یک سیستم پرسش و پاسخ تعاملی را از بین چند سیستم بدون نیاز به افراد خبره انتخاب نماید بتواند به راحتی یکی را برگزیند. بنابراین با توجه به توضیحات ارائه شده ارائه یک روش ارزیابی که بتواند در انتخاب یک سیستم پرسش و پاسخ تعاملی کمک نماید، امری ضروری است. شکل ۱-۲ نمایی از چگونگی انجام این کار را نمایش می‌دهد. برای دستیابی به یک مدل ارزیابی کامل، در نظر گرفتن تمامی جوانب از اهمیت بسزایی برخوردار است. معمولاً یک روش ارزیابی جامع باید شامل ۴ محور باشد:



شکل ۱-۲ نمای مربوط به فرآیند مدل‌سازی در رساله

۱- اقدامات سیستم محور<sup>۱</sup>: مانند کارایی یا عملکرد سیستم به عنوان مثال تعداد سؤالاتی که کاربر می‌پرسد تا به جواب خود دست پیدا نماید که این موضوع می‌تواند میزان بیانگر کارایی یک سیستم باشد.

۲- اقدامات فرآیند گرا<sup>۲</sup>: مانند نحوه عملکرد درونی هر قسمت از سیستم چه سخت‌افزاری و چه نرم‌افزاری مورد توجه باید قرار گیرد.

۳- اقدامات محصول گرا<sup>۳</sup>: کیفیت گزارشی که برای سیستم در نظر گرفته می‌شود در این مرحله مورد ارزیابی قرار می‌گیرد.

۴- اقدامات کاربر محور<sup>۴</sup>: منظور از این اقدامات سنجش رضایت، درک و اولویت کاربر می‌باشد.

با توجه به اینکه یک سیستم پرسش و پاسخ تعاملی از دو موجودیت سیستم و کاربر تشکیل شده و تعداد ویژگی‌ها از جنبه‌های مختلف، زیاد می‌باشند. امکان ایجاد یک روش استاندارد و جامع بسیار سخت است. بنابراین جایگزین کردن یک مدل به جای انسان تا حدود زیادی می‌تواند به فرآیند

<sup>1</sup> System-Oriented

<sup>2</sup> Process-Oriented

<sup>3</sup> Product-Oriented

<sup>4</sup> User-Oriented

خودکار کردن ارزیابی سیستم‌ها کمک نماید. با توجه به شرایط بیان شده در بالا، برای ارزیابی سیستم‌های پرسش و پاسخ تعاملی با یکسری موانع در این راه مواجه خواهیم بود که برای ساده‌تر شدن مسئله، نیاز به در نظر گرفتن یکسری مفروضات برای ادامه کارها خواهیم داشت که در ادامه به بیان آن‌ها خواهیم پرداخت.

## ۱-۴ چالش‌های موجود در فرآیند ارزیابی

ارزیابی سیستم‌های پرسش و پاسخ تعاملی یک فرآیند پرهزینه و وقت‌گیر است و هنوز محققین نتوانسته‌اند یک روش کلی برای ارزیابی سیستم‌های پرسش و پاسخ تعاملی ارائه نمایند تا بتوانند بر اساس آن سیستم معرفی شده را ارزیابی نمایند. برخی از چالش‌هایی که در مسیر ارزیابی سیستم‌های پرسش و پاسخ تعاملی با آن مواجه هستیم به شرح زیر است:

الف- ارائه یک معیار ارزیابی مناسب و مشترک برای ارزیابی سیستم‌ها

ب- انتخاب کردن ویژگی‌های مناسب از بین ویژگی‌های موجود جهت ارزیابی این سیستم‌ها از جنبه‌های مختلف

ج- تعیین تعداد ویژگی‌های بهینه که دارای دو شرط حداقل مقدار ویژگی و حداکثر کارایی ممکن (یعنی بیشترین دقت) باشد.

د- نحوه تعیین نوع تابع با توجه به اینکه ممکن است به یک تابع با  $n$  معادله (به ازای هر فرد شرکت کننده در ارزیابی یک معادله داریم) و  $m$  مجهول (مقادیر ویژگی‌های در هر معادله ارزیابی) برسیم. در این حالت استفاده از روش‌های معمول امکان‌پذیر نخواهد بود.

ه- کلاس‌بندی احتمالی مربوط به ویژگی‌ها و تعیین نوع ترکیب آن‌ها با یکدیگر به عبارت دیگر استفاده از یک مدل مشابه برای کلاس‌های در نظر گرفته شده یا مدل جدید برای ترکیب آن‌ها.

و- عدم وجود یک پایگاه داده‌ها مشترک برای ارزیابی سیستم‌های مختلف.

ز- ارائه یک روش کم هزینه و سریع.



ح- عدم دخالت انسان در فرآیند ارزیابی.

بنابراین در سال‌های اخیر روشی که بتواند به چالش‌های موجود پاسخ دهد مورد توجه محققین قرار گرفته است. از این‌رو در این رساله یک مدل آماری برای ارزیابی سیستم‌های پرسش و پاسخ تعاملی ارائه شده است تا بتواند گامی در جهت انجام هرچه بهتر شدن فرآیند ارزیابی سیستم‌های پرسش و پاسخ تعاملی برداشته شود. برای رسیدن به یک مدل مناسب ارزیابی، نیاز به در نظر گرفتن یکسری مفروضات است که بتوان ادامه راه را برای ما آسان‌تر نماید. تعدادی از آن‌ها به شرح زیر است:

الف- با توجه به انواع مختلف سیستم‌های پرسش و پاسخ تعاملی، سیستم در نظر گرفته شده برای ارزیابی، در حوزه سیستم‌های متنی قرار دارد که در این نوع سیستم‌ها کاربر با پرسیدن سئوالات متوالی به پاسخ مورد انتظار خود دست یابد.

ب- جواب‌های ارائه شده توسط سیستم تا حدی مختصر بوده و در قالب جمله و یا عبارت بیان می‌شود. به عبارت دیگر، سیستم در نظر گرفته شده نباید در پاسخ به سوال کاربر، یک قطعه متن یا سند ارائه نماید و در صورتی که سیستم چنین حالتی را برای کاربر در نظر گرفته باشد باید این قابلیت را داشته باشد تا خلاصه‌ای از سند در قالب جمله بیان گردد.

ج- ضریب همبستگی بین ویژگی‌های در نظر گرفته شده صفر فرض می‌گردد یعنی ویژگی‌ها مستقل از یکدیگر می‌باشند. در صورتی که وابستگی وجود داشته باشد فرض کرده که امکان کاهش بعد وجود دارد.

د- سیستم پرسش و پاسخ تعاملی در نظر گرفته شده، از کارایی مناسبی برخوردار باشد.

ه- معیار ارزیابی برای ضرایب در نظر گرفته شده در مدل آماری پیشنهادی، جذر میانگین مربعات خطا ( $RMSE^1$ ) است. همچنین ممکن است از معیار میانگین خطای مطلق درصد ( $MAPE^2$ ) که به عنوان انحراف درصد مطلق ( $MAPD^3$ ) نیز شناخته می‌شود، استفاده نمود.

<sup>1</sup> Root Mean Square Error

<sup>2</sup> Mean Absolute Percentage Error

<sup>3</sup> Mean Absolute Percentage Deviation

و- امکان تعریف معیارهای کیفی و تبدیل آن‌ها به مقادیر مناسب برای در نظر گرفتن در مدل پیشنهادی و همین‌طور بررسی عملگرهای ریاضی مجاز بر روی پارامترهای کیفی و تعداد سطح آن‌ها، نیز فراهم باشد.

ز- سیستم پیش فرض برای آزمون و اجرای ارزیابی مدل ارائه شده، سیستم طراحی شده در آزمایشگاه تکنولوژی زبان‌های طبیعی دانشگاه صنعتی شاهرود<sup>۱</sup>، می‌باشد.

با توجه به مفروضات بیان شده، اولین گام برای ارائه یک مدل آماری، شناسایی ویژگی‌های لازم در ارزیابی یک سیستم پرسش و پاسخ تعاملی و سپس به دست آوردن ضرایب هر یک از پارامترهای ارزیابی است. لذا در اولین گام نیاز هست تا ویژگی‌های شاخص را فهرست نموده، سپس به بررسی و محاسبه تأثیر هر یک از آن‌ها در ارزیابی سیستم‌های پرسش و پاسخ تعاملی بپردازیم.

## ۱-۵ فرضیه سیستم ارزیابی

با توجه به چالش‌های مطرح شده و مفروضات در نظر گرفته شده، برای رسیدن به یک مدل مناسب، چند نکته وجود دارد که باید قبل از مدل‌سازی به آن توجه نمود:

الف- چگونه می‌توان تصمیم گرفت که مدل پیشنهادی چند جمله‌ای است؟ آیا نیازمند تعریف شرایط برای رگرسیون چندگانه هستیم؟ آیا راه آسان‌تری برای خودکارسازی روند مربوط به مدل‌سازی وجود ندارد؟

ب- چگونه اطمینان حاصل نماییم که در انتخاب مدل، بیش برآزش اتفاق نمی‌افتد؟

ج- آیا مدل پیشنهادی در برابر نویز مقاوم است؟

د- آیا مدل پیشنهادی به راحتی قابلیت مقیاس پذیری به ابعاد بالاتر و یا مجموعه داده‌های بزرگتر را دارد؟

<sup>۱</sup> www.Hltech.ir

بنابراین برای انتخاب مدل با چالش‌های متعددی روبه‌رو هستیم. در فرضیه پیشنهادی، برای ارائه مدل، از رگرسیون بهره گرفته شد. فرض اولیه ترکیب ویژگی‌ها را با یک مدل خطی سری توانی با یکدیگر در نظر گرفتیم که رسیدن به معادله خط مناسب، خود یکی دیگر از چالش‌های موجود در سر راه معرفی مدل می‌باشد. مدل اولیه که برای ارزیابی سیستم پرسش و پاسخ تعاملی در نظر گرفته ایم به شکل معادله ۱-۱ است.

$$y = A_1 X_1^{T_1} + A_2 X_2^{T_2} + \dots + A_{n-1} X_{n-1}^{T_{n-1}} + A_n X_n^{T_n} = \sum_{i=1}^n A_i X_i^{T_i} \quad (1-1)$$

که در آن ضریب  $A_i$  بیانگر میزان تأثیر هر یک از این ویژگی‌ها ( $X_i$  ها) و  $T_i$  مرتبه هر یک از این ویژگی‌ها را بیان می‌کند. به دست آوردن هر یک از این ضریب‌ها و تعیین مقادیر آن‌ها یکی از کارهای اساسی صورت پذیرفته در این رساله است. در این مدل، به عنوان فرض اولیه، همبستگی بین این ضرایب را برابر صفر در نظر گرفته شد. در نهایت برای به دست آوردن مدل مناسب رگرسیون‌های متفاوت مانند رگرسیون خطی چند متغیره<sup>۱</sup>، غیرخطی و توانی استفاده و در نهایت با استفاده از برنامه‌نویسی بیان ژن<sup>۲</sup> توانستیم به بهترین مدل دست پیدا نماییم. برای آشنایی بیشتر خواننده، ضمیمه الف و ب به ترتیب به بررسی رگرسیون و روش برنامه‌نویسی بیان ژن پرداخته است.

<sup>1</sup> Multi Linear Regression

<sup>2</sup> Gene Expression Programming



## فصل دوم - سابقه تحقیق

بررسی روند سیستم‌های پرسش و پاسخ

بررسی روند سیستم‌های پرسش و پاسخ تعاملی

روش‌های ارزیابی سیستم‌های پرسش و پاسخ تعاملی

همانند بسیاری از زمینه‌های مربوط به سیستم‌های پرسش و پاسخ تعاملی، ارزیابی نقش مهمی در این سیستم‌ها ایفا می‌نماید. علی‌رغم این موضوع، تقریباً در زمینه ارزیابی سیستم‌های پرسش و پاسخ تعاملی روش خاصی وجود ندارد که به ارزیابی کلی این سیستم‌ها پرداخته باشد و تنها روش‌های موجود، روش‌هایی هستند که با روش ارزیابی به کار برده شده در QA و سیستم‌های دیالوگ انطباق دارند [۴]. مشکل اصلی طراحی روش ارزیابی در سیستم‌های IQA به دلیل عدم امکان پیش‌گویی بخش تعامل این سیستم‌ها است. به همین منظور، باید انسان در فرآیند ارزیابی دخالت داشته تا فرآیند پردازش در زمان و هزینه کمتر و به‌صورت ساده‌تر صورت پذیرد [۵]. در این فصل مروری بر روند پیدایش سیستم‌های پرسش و پاسخ، کارهای صورت گرفته در زمینه سیستم‌های پرسش و پاسخ تعاملی و روش‌های موجود ارزیابی برای این سیستم‌ها می‌پردازیم.

## ۲-۱ روند پیدایش سیستم‌های پرسش و پاسخ

سیستم‌های QA اولیه همانند Baseball [۶] و Lunar [۷] در دهه ۶۰ میلادی، جزء اولین سیستم‌ها با پایگاه داده ساختارمند بودند [۱]. سؤالات ارائه شده به این سیستم‌ها عموماً از طریق روش‌های NLP تحلیل می‌شدند که بعدها جهت ایجاد یک پرس و جوی استاندارد در پایگاه داده مورد استفاده قرار گرفتند. سیستم‌های محاوره‌ای نظیر GUS [۸] و ELIZA [۹] نیز از یک پایگاه داده ساختارمند به عنوان منبع علمی استفاده می‌کردند. عامل اصلی در محدودیت این سیستم‌ها، اطلاعات ذخیره شده در پایگاه داده آن‌ها بود که فقط امکان پاسخگویی به پرسش‌ها، در زمینه‌های محدودی را فراهم می‌نمود. با این حال، در پژوهش‌های اخیر، محدودیت‌های سیستم‌های دانش محور پذیرفته شده‌اند. کلارک و همکارانش راهکاری برای افزایش متن برخط (راهنمای پویا) به همراه قابلیت پاسخ به سؤالات دانش محور، ارائه نمودند. راهکار ترکیبی آن‌ها، این امکان را برای کاربران فراهم نموده بود که به پاسخ سؤالات متداول و سؤالات غیرقابل پیش‌بینی دسترسی داشته باشند. آن‌ها این ویژگی را از طریق موتور استنتاج به سیستم اضافه کرده بودند [۱۰]. در کار مشابه دیگری سیستم Quarc و

Cqarc که به ترتیب توسط ریلاف [۱۱] و هااو [۱۲] توسعه یافتند، از قوانین اکتشافی در سؤالات (جهت تشخیص کلاس پرسش)، بهره گرفته بودند. طبق نظر آن‌ها، رده‌بندی کلاس پرسش می‌تواند از یک سیستم به سیستم دیگر متفاوت باشد. بعضی از سیستم‌ها از رده‌بندی عمومی برای کلاس‌های معنایی همانند (who, when, what, where, why) بهره می‌برند درحالی‌که برخی دیگر از رده‌بندی حوزه معین استفاده می‌کردند. حجازی و همکارانش در سال ۱۳۸۳ یک سیستم پرسش و پاسخ مبتنی بر هستان‌شناسی برای پاسخگویی به پرسش‌های حوزه تخصصی مخابرات فیبرنوری، با قابلیت استخراج و دسته‌بندی خودکار مستندات ارائه نمودند. این سیستم، پرسش‌های کاربر را به زبان انگلیسی دریافت کرده و به کمک استدلال روی گراف هستان‌شناسی پاسخ دقیق را استخراج و به همراه پاراگراف‌های خلاصه‌سازی شده مرتبط در اختیار کاربر قرار می‌داد. آن‌ها نشان دادند که با تغییر در ساختار هستان‌شناسی این سیستم، می‌توان دامنه پاسخگویی را تغییر داد. نتایج به دست آمده از ۱۰۰ پرسش در دامنه فناوری مخابرات، نشان داد که این سیستم از دقت و سرعت قابل قبولی برخوردار است. اما باید توجه داشت که این سیستم در دامنه محدودی از دانش کار می‌نمود [۱۳].

یوسفی در سال ۱۳۹۱ سیستم پرسش و پاسخی در حوزه اطلاعات پرواز با دامنه محدود ارائه نمود که توانایی پاسخگویی به پرسش‌ها را دارا بود. این سیستم با توجه به معماری ارائه شده، پردازش‌هایی را بر روی پرسش انجام می‌داد و با ایجاد پرس و جویی، آن را بر روی پایگاه داده اجرا نموده و جواب ایجاد شده را در یک یا چند سطر به کاربر نشان می‌داد [۱۴]. ساجدی و خانی اولین سیستم پرسش و پاسخ فارسی دامنه نامحدود و وب مبنا را به نام جويا معرفی نمودند. به دلیل نبود مجموعه داده ارزیابی در زبان فارسی، آن‌ها یک مجموعه داده ارزیابی برای این سیستم تهیه نمودند که این مجموعه داده شامل ۴۱۲ پرسش متنوع و پاسخ متناظر آن است. سیستم پیشنهادی به ۸۰ درصد صحت دست پیدا کرده است. جويا همانند سیستم‌های استارت و واتسون دامنه نامحدود بوده و برخلاف واتسون اطلاعات خود را از وب اخذ می‌کند. البته در نسخه اول جويا، دانش مبتنی بر ویکی‌پدیای فارسی

است. جویا همانند سیستم‌های استارت و Wolfram-alpha سؤال‌های حقیقی را پاسخ می‌دهد و نمی‌تواند مانند سیستم واتسون، پرسش‌های پیچیده را پاسخگو باشد [۱۵].

سیستم IBM یکی از سیستم‌های پیشرو در زمینه استفاده از مدل‌های آماری است. این سیستم، برای پرس و جوی مبتنی بر خصوصیات کلمات، از مدل آنتروپی ماکزیمم<sup>۱</sup> و یا N-gram ها استفاده می‌نماید [۱۶]. از دیگر سیستم‌های مطرح شده، سیستم Rocchio است که توسط Moschitti ایجاد گردید. در این سیستم برای طبقه‌بندی پرسش از SVM استفاده گردید. آن‌ها مدلشان را بر روی Reuters- 21578 تست نمودند و برای محاسبه تطابق امتیاز دریافتی بین پرسش و متن، مدل سنجش تشابه را بکار بستند [۱۷]. Zhang و همکارانش یک سیستم پرسش و پاسخ چینی را توسعه دادند که از کلاس بندی کننده SVM استفاده می‌کرد [۱۸]. Quarteroni و همکارانش سیستم QA را پیشنهاد کردند که جهت طبقه‌بندی پرسش‌ها از کلاس بند SVM استفاده می‌نمود [۱۹]. Berger و همکارانش به کارگیری روش‌های استاتیکی جهت یافتن پاسخ در سیستم‌های QA را مورد بررسی قرار دادند و دریافته‌اند که این تکنیک‌ها برحسب ویژگی‌هایی نظیر اندازه واژگان مجموعه داده اصلی، هم‌پوشانی بین پرسش و پاسخ و هم‌پوشانی بین چندین پاسخ بسیار خوب عمل می‌نماید. آن‌ها از تکنیک‌های استاتیکی نظیر داده کاوی N-gram، مدل تشابه و روش تشابه Okapi جهت پروسه یافتن پاسخ در سیستم QA استفاده کردند [۲۰]. کای و همکارانش یک مدل تشابه، جهت محاسبه تشابه بین پرسش و پاسخ ارائه نمودند. این مدل بر اساس خصوصیات متفاوت شامل، تشابه در کلمه کلیدی، طول، ترتیب، فاصله کلمه کلیدی بکار رفته در پرسش و پاسخ معرفی شده بود [۲۱]. گوناوردونا و همکارانش به بررسی فن‌های ارزیابی برای سیستم‌های QA پرداختند. آن‌ها در مقاله خود، از یک روش تطبیق الگو برای انجام این کار استفاده کردند [۲۲]. از کارهای ارائه شده در بالا می‌توان دریافت که در طی سال‌های پژوهش و فعالیت در حوزه ایجاد سیستم‌های QA، تعداد بسیار زیادی از نمونه‌های دانشگاهی و صنعتی پیاده‌سازی و طراحی شدند که در آغاز اکثر سیستم‌ها تنها پاسخگوی

---

<sup>1</sup> Entropy Maximium



پرسش‌هایی بودند که پیرامون موضوع خاصی بودند ولی بعد از سال ۱۹۹۹ سیر تکامل سیستم‌های QA بدین صورت پیش رفت که دامنه متنوعی از موضوعات را پوشش دهند. جدول ۱-۳ بخشی از سیستم‌های مطرح در طی سالیان گذشته را نمایش می‌دهد.

## ۲-۲ مروری بر کارهای صورت پذیرفته در سیستم‌های پرسش و پاسخ تعاملی

کارهای متعددی در زمینه طراحی سیستم‌های پرسش و پاسخ تعاملی صورت پذیرفته است. در طراحی این سیستم‌ها، هدف بدین صورت است که کاربر به جای پرسیدن یک سؤال پیچیده برای رسیدن به اطلاعات مورد نظر خود، چند سؤال ساده را مطرح نماید. بعلاوه اطلاعات جست‌وجو شده به روشی تکراری و با تکرار سؤال صورت پذیرد. از جمله این سیستم‌ها می‌توان به سیستم VAN [۲۳] اشاره کرد که از این فرض بهره گرفته بود. همچنین سیستم‌های IQA ای وجود دارند که به طور خودکار بعضی از عبارات کوئری را بسط داده و یا حذف نموده و همیشه پتانسیل پاسخ به سؤال را دارند. یکی از کارهای متداول در این نوع از سیستم‌ها این است که از کاربر خواسته می‌شود که سؤالاتش را به صورت واضح و شفاف پرسیده و هنگامی که سیستم چندین جواب را نمایش می‌دهد، کاربر با استفاده از این اطلاعات، سؤال را باید بازنویسی نمایند. همان‌طور که قبلاً اشاره شد سیستم‌های IQA در سه دسته مختلف قرار می‌گیرند که در ادامه تحقیقات موجود در این سه دسته آورده شده است. کی‌یو و گرین سیستمی را طراحی کردند که می‌توانست با غلبه بر درخواست‌های محدود شده در حیطه عبارات، به ارتقاء سیستم‌های دیالوگ کمک قابل توجهی نماید. سیستم طراحی شده، بدین صورت عمل می‌نمود که سیستم دیالوگ می‌بایستی (۱) به جایگذاری ساختار کوئری با ساختار پیشنهاد شده، (۲) کشف سریع درخواست‌های اطلاعات خارج از سطح محدوده و (۳) ارائه پاسخ‌های مشارکتی می‌پرداخت [۲۴]. وارگاس روشی را پیشنهاد کرد که یک سیستم دیالوگ می‌توانست نتایج کوئری‌های پایگاه داده را مدیریت نماید. هدف آن‌ها از این کار، ارائه روش‌های مؤثر مدیریت دیالوگ برای عرضه اطلاعات به کاربران بود.

جدول ۱-۳ روند پیشرفت سیستم‌های QA

نام سیستم	نام مرجع	سال	دامنه سیستم	
			باز	محدود
Baseball	Baseball: an automatic question-answerer	۱۹۶۱	*	
LUNAR	Building a question answering test collection	۱۹۷۲	*	
SHRDLU	Wikipedia ( <a href="http://en.wikipedia.org/wiki/SHRDLU">http://en.wikipedia.org/wiki/SHRDLU</a> )	۱۹۷۲	*	
MYCIN	Wikipedia ( <a href="http://en.wikipedia.org/wiki/MYCIN">http://en.wikipedia.org/wiki/MYCIN</a> )	۱۹۷۵	*	
AskMSR	An Analysis of the AskMSR Question-Answering System	۲۰۰۲	*	
AnswerBus	AnswerBus Question Answering System	۲۰۰۲	*	
START	Viewing the web as a virtual database for question answering	۲۰۰۴	*	
Evi (True Knowledge)	Question Classification using semantic, syntactic and lexical features ( <a href="http://en.wikipedia.org/wiki/Evi_(software)">http://en.wikipedia.org/wiki/Evi_(software)</a> )	۲۰۰۷	*	
LogAnswer (used German wikipedia)	Question Classification using semantic, syntactic and lexical features	۲۰۰۹	*	
EAGLEi	Question Answering for Biology and Medicine	۲۰۰۹	*	
WolframAlpha	A survey on question classification techniques for question answering	۲۰۰۹	*	
IBM Watson	Wikipedia ( <a href="http://en.wikipedia.org/wiki/IBM_Watson">http://en.wikipedia.org/wiki/IBM_Watson</a> )	۲۰۱۰	*	
QARAB	A survey on question answering technology from an information retrieval perspective	۲۰۱۱	*	
سیستم پرسش و پاسخ مبتنی بر هستان شناسی برای حوزه مخابرات	سیستم پرسش و پاسخ مبتنی بر هستان شناسی برای حوزه مخابرات با قابلیت استخراج و دسته‌بندی خودکار	۱۳۸۵	*	
جویا	جویا: یک سیستم پرسش و پاسخ فارسی	۱۳۹۵	*	

بنابراین روش پیشنهادی توسط آن‌ها نمی‌توانست برای حالاتی که نیاز به اطلاعات بیشتری بود، راه‌حلی ارائه نماید و یا درباره بعضی موارد نمی‌توانستند خطا را نادیده بگیرند. به‌طور کلی بخش عمده تحقیق محققان در این زمینه بر روی چگونگی دستیابی به محدودیت‌های کوئری معتبر، چگونگی ارائه نتایج برای کاربر و زمان ارائه نتایج متمرکز شده است [۲۵]. محققین این حوزه سعی داشتند که مسئله تبادل بین طول دیالوگ (چه تعداد سؤال پرسیده شود) و تعداد آیتم‌های موجود در لیست نهایی را مطرح نمایند و سیستم‌هایشان بهترین استراتژی‌های دیالوگ را برای پاسخ تعاملی به سؤال فراهم کنند. در این گونه از سیستم‌ها، برای ارزیابی کلی سیستم، از مقایسه بهترین نتایج به دست آمده با آستانه دستی بهره می‌گرفتند. ماگینی برای ارائه و توسعه سیستم IQA، بر روی سیستم ارتقاء یافته و پیشرفته QA تکیه داشت. بر خلاف روش‌های توصیف شده در بالا، وی از سیستم قدیمی QA به عنوان نقطه شروع استفاده کرد و روش‌های تعامل کاربردی در موارد عدم دستیابی به موفقیت را پیشنهاد کرد [۲۶]. دورنسکو و همکارانش راه‌حلی را عنوان کردند که اطلاعات مربوط به نوع پاسخ پیش‌بینی شده و محدودیت‌های مربوط به سؤال را تحلیل می‌نمودند و در صورت عملکرد ناموفق سیستم در پاسخ به سوال، از یک ابر داده، برای آغاز روند تعامل با کاربری که سؤال را مطرح کرده بود، استفاده می‌نمودند و با کمک آن محدودیت‌های مربوط به اصلاح و تغییر سؤال را اعمال می‌نمودند. راه‌حل پیشنهاد شده در این سیستم، همانند راه‌حل‌های ارائه شده در سیستم‌های قبلی بود [۲۷]. کوارترونی و همکارانش سیستم QA پیشنهاد کردند که ترکیبی از یک سیستم QA دامنه باز و چت بات<sup>۱</sup> بود. سیستم ارائه شده برای دستیابی به پاسخ مربوط به سؤالات، از اینترنت بهره می‌جست. آن‌ها یک AIML<sup>۲</sup> را به سیستم تعاملی اضافه کردند که آخرین گفتار کاربر در برابر طیف گسترده‌ای از الگوهای دیالوگ انطباق می‌یافت و سپس پاسخ همسان و منسجمی منطبق با چنین الگویی تولید می‌شد. ارزیابی‌های چندگانه‌ای بر روی سیستم پیشنهادی انجام شد و نشان داد که بسیاری از کاربران (ارزیابی اول ۸۷/۵٪ و ارزیابی دوم ۵۸/۳٪) سیستم تعاملی را به جای سیستم

<sup>۱</sup> Chatbot

<sup>۲</sup> Artificial Intelligence Markup Language

پرسش و پاسخ ساده در اولویت قرار داده اند [۱۹]. لئو مزایای کاربرد IQA در رابطه با تشخیص عیوب سخت افزاری کامپیوتر را مورد بحث و بررسی قرار داد. در این مقاله، افزایش دقت از ۱۳٪ برای پاسخگویی به سؤالات ساده تا ۸۱ درصد برای پاسخگویی به پرسش و پاسخ تعاملی گزارش شد [۲۸].

تانگ سیستم پیشنهادی خود را با توجه به موقعیت سؤال ها، مبهم یا نامشخص بودن آن ها ارائه نمود. در این سیستم سؤالات اضافی، برای مطلع ساختن کاربر و رفع ابهام، مطرح می گردید. آن ها از وردنت چینی برای ایجاد این سؤالات اضافی استفاده کرده بودند. در این مقاله، دقت پیشنهاد ها و موارد دیگر گزارش شده بود، با این وجود، تأثیر کلی QA ارزیابی نشده بود [۲۹]. شهرآیینی هدف از تحقیق خود را طراحی یک سیستم پرسش و پاسخ مبتنی بر تعامل با استفاده از تکنیک های آماری جهت استخراج دانش نهفته در متون ساختار نیافته بیان کرده است. وی در تحقیق خود سیستمی برخلاف سیستم های پیشین که عمدتاً از تجزیه و تحلیل معنایی و گرامری استفاده می کردند، ارائه نمود که مستقل از زبان بوده و از روش های مبتنی بر معنا و گرامر استفاده نمی کند. بنابراین با در اختیار داشتن پایگاه دادگان مناسب هر زبان، سیستم طراحی شده می تواند به سؤالات مطرح شده به آن زبان پاسخ دهد. در طراحی سیستم پیشنهادی ایشان، از پایگاه دادگان فارسی که به این منظور ایجاد گردیده بود، جهت آموزش و ارزیابی استفاده شده است. میزان کارایی سیستم با استفاده از معیارهای کمی و کیفی سنجیده شده است. مقایسه مقادیر به دست آمده، حاکی از افزایش ۲۲/۳٪ دقت پاسخ گویی نسبت به سیستم های پرسش و پاسخ را نشان داد. همچنین بررسی نظرات داده شده توسط کاربران نشان از رضایت آن ها از کیفیت تعامل برقرار شده با سیستم است [۱].

اسکوتن دسته بندی سؤالات متوالی را مورد بررسی قرار داد و چالش هایی که در این خصوص وجود داشتند را عنوان نمود [۲۳]. آن ها چگونگی عملکرد این روش و تأثیرگذاری متقابل آن را دو نمونه سیستم IMIX [۳۰] و ریتل [۳۱] مورد بررسی قرار دادند. هاراباگوئی در سیستم پیشنهادی خود سعی در پیشگویی دامنه سؤالات کاربر با پرسیدن موضوع و مفاهیم مربوط به آن را داشت. در این سیستم، بعد از مطرح شدن سؤال توسط کاربر، این امکان توسط یک سیستم تعاملی ایجاد می گردید که کاربر

نه تنها می توانست به سوال پاسخ دهد، بلکه می توانست ارائه سؤالات متوالی را نیز پیشنهاد نماید. آن ها استدلال کردند که توانایی یک سیستم های تعاملی به سه عامل تأثیر بازنمایی موضوع گفتگو، شناسایی پویای یک گفتگو و توانایی بازگرداندن پاسخ های مرتبط به یک سؤال خاص بستگی دارد. نتایج ارائه شده توسط آن ها حاصل از کار کردن با یک سیستم تعاملی به نام FERRET بود که تحت نظارت برنامه AQUAINT مربوط به سازمان ARDA قرار داشت و برای ۸ سناریو مختلف و بیش از ۳۰ کاربر بکار گرفته شده بود. همچنین آن ها نشان دادند که کیفیت تعامل تولید شده به وسیله یک سیستم IQA، می تواند به وسیله پیش بینی رنج سؤالات ممکن در مورد موضوع گفتگو، ارتقاء یابد. بنابراین اگر از سؤالات ممکن در مورد موضوع گفتگو، یک پایگاه داده موجود باشد، دقت سیستم بالاتر خواهد رفت. انتخاب سؤالات از پایگاه داده بر اساس معیار شباهت در رتبه سؤال، برای هر سؤال کاربر ارائه می شد که آن ها برای محاسبه معیار شباهت از هفت روش مختلف بهره گرفتند. در نهایت آن ها سیستم FERRET را به وسیله اندازه گیری کارایی، مؤثر بودن و رضایت مشتری ارزیابی نمودند. ارزیابی انجام شده حاکی از این بود که پایگاه داده در نظر گرفته شده به ارتقاء کارایی و کارآمدی سیستم کمک می کند و به طور کلی به ارتقاء سطح رضایت کاربر نیز کمک شایانی می نماید [۲۸]. برناردی و کریسکنر در مقالات ارائه شده توسط آن ها [۳۲-۳۴]، روش پاسخگویی به سؤالات متوالی را مورد بررسی قرار دادند. آن ها بر این باور بودند که یک سیستم IQA با متمرکز بودن بر روی پیش گویی سؤالات متوالی، می تواند قابلیت سیستم را ارتقا دهد. آن ها آنچه را که منجر به انسجام یک دیالوگ می شد و همچنین آنچه را که کاربر بر روی آن متمرکز است، را به دقت مورد بررسی قرار دادند. آن ها یک مازول ردیابی موارد مهم، در دیالوگ ها را آزمایش کردند و نتایج ارائه شده، حاکی از تأثیر آن در ارتقاء عملکرد سیستم IQA بود. این محققین در تحقیق دیگری فقط به بررسی سؤالات مرتبط با موضوع نپرداخته بلکه یک گام فراتر رفته و به بررسی موضوعات قبلی در دیالوگ پرداخته بودند. آن ها، روابط متفاوتی را شناسایی کردند که بین سؤالات متوالی و پردازش دیالوگ قرار داشتند و می توانست به انتخاب جواب صحیح در بین جواب های کاندید شده کمک نماید. با بکار بردن نتایج قبلی،

کریسکندر و همکارانش روش‌هایی را پیشنهاد کردند که بین انتقال موضوع و تداوم موضوع قبلی تمایز ایجاد می‌کرد و نتایج سیستم IQA را ارتقا می‌داد. آن‌ها به این نکته اذعان داشتند که تصمیم‌گیری در خصوص کاربرد روش‌ها و بازیابی پاسخ، بسیار حائز اهمیت است. همچنین آن‌ها به بررسی ویژگی‌های سطحی مثل شباهت لغوی، شباهت توزیعی، شباهت معنایی و توالی عملکردی تکیه داشتند. اگر چه تنها در کاربرد ویژگی‌های سطحی نتایج نوید دهنده‌ای گزارش نمودند. برناردی و همکارانش در تحقیق دیگر خود با افزودن ویژگی‌های عمیق، مدل قبلی خود را ارتقاء دادند. با وجود این، نتایج حاکی از این هستند که این ویژگی‌ها فراتر از ویژگی‌های سطحی هستند، اما تلفیق ویژگی‌های عمیق همراه با افزایش دقت سیستم تعاملی پاسخگویی به سؤال خواهد بود. اسکوتن واکر IQA را به عنوان سیستم QA محتوای گفتار متوالی برای کاربران معرفی کردند. زمانی که کاربران با پاسخ مناسب مواجه نمی‌شدند می‌توانستند برای رسیدن به پاسخ مناسب شروع به پرسیدن سؤال نمایند. هدف اصلی این سیستم، تصمیم‌گیری در خصوص نوع سؤالات متوالی بود که با آن مواجه شده و اینکه از چه استراتژی باید استفاده می‌شد. بررسی‌های انجام شده در مورد متن و بافت سؤال با بررسی‌های انجام شده در خصوص سؤالات متوالی، حاکی از شباهت آن‌ها با یکدیگر بود [۳۵]. مقالات اخیر مانند وانگ [۳۶]، والتینگر [۳۷] روش‌های مربوط به دسته‌بندی نوع سؤال را مورد بررسی قرار داده‌اند. والتینگر مسئله تمرکز بر روی موضوع سؤال را مورد بررسی قرار داد. آن‌ها سعی داشتند کانون اصلی سؤال و همچنین اطلاعاتی که به تشخیص سؤال کمک می‌نماید، را شناسایی کنند. آن‌ها معتقد بودند که شناسایی اطلاعات اضافی متنی و سؤالات مرتبط، برای سیستم پاسخگویی به سؤال، مفید است و می‌توان با بکار بردن آن‌ها در حیطه سناریوهای تعاملی، زمان و دقت پاسخگویی را ارتقاء داد.

## ۲-۳ روش‌های ارزیابی سیستم‌های پرسش و پاسخ تعاملی

ارزیابی سیستم‌های پرسش و پاسخ تعاملی به منظور تعیین و ارتقای کارایی آن‌ها از اهمیت زیادی برخوردار است. با این وجود هنوز روش استاندارد و مخصوصی برای ارزیابی این سیستم‌ها ارائه نشده

است و ارزیابی معمولاً با استفاده از روش‌های ارزیابی سیستم‌های پرسش و پاسخ خودکار و نیز سیستم‌های مکالمه محور صورت می‌گیرد. در ارزیابی سیستم‌های پرسش و پاسخ تعاملی علاوه بر ارزیابی کمی از ارزیابی کیفی نیز استفاده می‌شود که نیازمند مشارکت کاربران در فرآیند ارزیابی برای تعیین میزان موفقیت تعامل بین سیستم و کاربر است. در ادامه روش‌های ارزیابی کمی و کیفی که در سیستم‌های پرسش و پاسخ تعاملی مورد استفاده قرار می‌گیرند، تشریح می‌شوند. کارهای صورت گرفته قابل توجهی در ارزیابی سیستم‌های پرسش و پاسخ در زمینه استفاده از کاربران واقعی و یا مشکلات واقعی وجود دارد. ارزیابی سیستم‌های QA بسته به ارزیابی سؤالات پیچیده یا ساده (مثل تعریف، روابط و سناریوهای مربوط به سؤالات) متفاوت است. ارزیابی کارایی سیستم‌های پرسش و پاسخ از نظر صحت و دقت پاسخ‌گویی آن‌ها، با توجه به نوع سیستم، اندکی متفاوت است.

یکی از روش‌های ارزیابی مورد استفاده در سیستم‌های پرسش و پاسخ ایجاد یک مجموعه‌ای از سؤالات و پاسخ به نام «مجموعه استاندارد طلایی» و سپس با استفاده از میزان منطبق بودن سیستم با این مجموعه استاندارد طلایی، توانایی یک سیستم سنجیده شود. البته این روش برای سؤالات پیچیده و مبهم هنوز تقویت نشده است [۵]. ضمن اینکه در بعضی از این سیستم‌ها، پاسخ صحیح یک سؤال می‌تواند به شکل‌های مختلفی بیان شود. سؤالات ممکن است بیش از یک پاسخ صحیح داشته باشند. ارزیابی این گونه سیستم‌ها معمولاً به صورت دستی و توسط شخص صورت می‌گیرد. مسائل طراحی کلی برای ارزیابی این سیستم‌ها توسط Tague [۳۸] در سال ۱۹۹۲ بیان شد. برای نمونه می‌توان به ارزیابی سیستم‌های پرسش و پاسخ شرکت‌کننده در اجلاس سالانه بازیابی متون اشاره نمود. بیشتر ارزیابی سیستم‌های QA در این اجلاس انجام شده است که این ارزیابی‌ها بیشتر به جای اینکه مبتنی بر یک سیستم باشد بر اساس کاربر صورت پذیرفته است. با برگزاری این اجلاس، دوره جدیدی در طراحی و ارزیابی سیستم‌های پرسش و پاسخ آغاز گردید [۳۹، ۴۰]. در حقیقت بیشتر کارهای صورت گرفته ارزیابی در زمینه استخراج پاسخ و نحوه تعامل و استفاده از آن در این اجلاس انجام شده است. اجلاس TREC-8 با اعلام رویکرد جدید خود، طراحان سیستم‌های پرسش و پاسخ را

به طراحی سیستم‌هایی تشویق نمود که با بهره‌گیری از مجموعه بزرگی از اسناد متنی در موضوعات مختلف، بتوانند پاسخ کوتاهی برای سؤال کاربر که درباره موضوعات مختلف با زبان طبیعی مطرح شده، ارائه نمایند که بعدها این‌گونه سیستم‌ها را سیستم‌های با دامنه نامحدود<sup>۱</sup> و گاهی سیستم‌های پویا نامیدند. به‌طور کلی در سیستم‌های ارائه شده از یک ارزیاب انسانی بهره گرفته می‌شد. البته این نکته لازم به ذکر است که این ارزیاب‌ها قبلاً باید آموزش دیده باشند. روش‌های مطرح شده در این زمینه، بیشتر برای کنترل صحت و درستی جواب‌های برگشت داده شده به کاربران بود و در حیطه پاسخ‌دهی سیستم‌های IQA بسیار حائز اهمیت است. با وجود این، روش‌ها نمی‌توانند اطلاعات کافی درباره کیفیت سیستم IQA فراهم نمایند. به همین دلیل، روش‌های ارزیابی مربوط به سیستم‌های دیالوگ بیشتر مورد استفاده قرار می‌گیرند. با توجه به ماهیت سیستم و نحوه بازگرداندن پاسخ، ارزیابی کمی بر اساس یک یا چند معیار ارزیابی انجام می‌شود.

هدف از ارزیابی کمی سیستم‌های پرسش و پاسخ تعاملی، تعیین میزان صحت پاسخ بازگردانده شده توسط این سیستم‌ها است. این ارزیابی، اطلاعات کافی درباره کیفیت تعامل سیستم با کاربر و اینکه آیا تعامل موفقیت‌آمیز خاتمه یافته است یا خیر، ارائه نمی‌کند. به همین دلیل علاوه بر ارزیابی کمی، این سیستم‌ها از نظر کیفی نیز مورد سنجش و ارزیابی قرار می‌گیرند تا کیفیت تعامل سیستم با کاربر و میزان رضایتمندی آن تعیین شود [۴]. جهت ارزیابی کیفیت تعامل، پرسشنامه‌ای تهیه شده و در اختیار کاربران قرار می‌گیرد تا با تکمیل آن میزان رضایت خود را از کیفیت تعامل اعلام کنند. این پرسشنامه‌ها طوری طراحی می‌گردد که بتوانند بر سه ویژگی اساسی کارایی<sup>۲</sup>، اثربخشی<sup>۳</sup> و رضایت مشتری سیستم‌های پرسش و پاسخ تأکید داشته باشد. به‌طور کلی، پاسخ صحیح در یک سیستم پرسش و پاسخ تعاملی و ارزیابی جواب‌های ممکن، با روش‌های متفاوتی بیان می‌شود. اکثر سیستم‌های موجود در ارزیابی، از ارزیابی انسان بهره می‌گیرند.

<sup>۱</sup> Open Domain Question Answering systems (ODQA)

<sup>۲</sup> Efficiency

<sup>۳</sup> Effectiveness



اسپیتر، آخرین ویژگی‌های مربوط به یک پرسش ساده و کامل کردن پرسشنامه را مورد بررسی قرار داده است و کیفیت سیستم را بر اساس تکمیل شدن پرسشنامه درجه‌بندی می‌نماید. سؤالات در نظر گرفته‌شده، شامل درخواست ارزیابی سیستم است، اما چنین روشی خود قابل بحث و بررسی است [۴۱]. هاراباگوئی عبارات رد و بدل شده در دیالوگ یک سیستم IQA را مورد ارزیابی قرار داده است. کارایی مؤثر سیستم وابسته به تعداد سؤالات تعریف شده که کاربر برای یافتن اطلاعات خاص مطرح می‌نماید، تعریف شده است [۴۲]. دورنسیکو و اوراسان در روشی که ارائه نمودند، از ارزیاب انسانی، درخواست کردند تا سؤالات متوالی ایجاد شده توسط سیستم را در مواردی که سیستم برای سؤال جواب مناسبی ندارد به دقت رتبه‌بندی نماید. آن‌ها با این کار سعی داشتند با اعمال محدودیت‌هایی برای اصلاح شرایط و به منظور دستیابی کاربران به پاسخ مناسب، فرآیند را طوری به جلو هدایت نمایند که پاسخدهی به آن‌ها در اولویت قرار گیرد [۲۷]. کلی چندین مقاله در زمینه ارزیابی ارائه نموده است [۴۳]. وی و همکارانش در سال ۲۰۰۷ مقاله‌ای [۴۴] ارائه کردند که در آن به ارزیابی عملکرد چهار سیستم IQA با کاربر واقعی پرداخته شده است. در مقاله اول ایشان هدف از ارائه آن شناسایی پتانسیل معیارهای ارزیابی برای سیستم‌های IQA با تجزیه و تحلیل نظرات ارزیابی ساخته شده توسط کاربران برای چنین سیستم‌هایی بود. آن‌ها داده‌های کیفی از تحلیلگران اطلاعاتی در طول مصاحبه‌ها را جمع‌آوری و از گروه‌هایی متمرکز برای شناسایی موضوعات مشترک مربوط به عملکرد، استفاده و قابلیت استفاده مجدد، برای آنالیز این داده‌ها، بهره جستند. این داده‌ها به عنوان بخشی از یک کارگاه سه روزه ارزیابی فشرده از یک سیستم تعاملی به نام HITIQA جمع‌آوری شد. برای شناسایی و دسته‌بندی‌های داده‌های جمع‌آوری شده از برنامه‌نویسی خاصی بهره گرفته بودند. نتایج نشان می‌داد که پتانسیل معیارهای ارزیابی برای سیستم‌های IQA، می‌تواند به عنوان راهنمای توسعه، طراحی و ارزیابی سیستم‌های آتی استفاده شود. در این مقاله، آن‌ها بر روی تحلیل سیستم QA تمرکز کردند. ارزیابی‌های صورت گرفته برای سیستم‌های QA معمولاً بر روی چگونگی یافتن پاسخ یک سؤال تمرکز می‌کنند تا به بررسی این‌که چه اتفاقی رخ می‌دهد تا یک سیستم بتواند ارتباط لازم بین

اطلاعات را برای دستیابی به پاسخ برقرار نماید. در صورتی که در سیستم‌های تعاملی بیشتر خواستار زمان، توجه و تلاش کاربر هستند. از آنجا که این روند خود، بخشی از محصول است. بنابراین، این موضوع سبب گردید تا کلی و همکارانش به ارزیابی کاربر که محوری حیاتی در ارزیابی‌ها هستند بپردازند. آن‌ها دو کارگاه ارزیابی سه روزه HITIQA با تحلیلگران اطلاعاتی به عنوان شرکت‌کنندگان در این مطالعه برگزار کردند و تعدادی تکنیک جمع‌آوری داده‌ها از جمله مشاهده، ورود به سیستم، پرسشنامه و مصاحبه را بکار بستند. این کارگاه‌ها در یک موسسه داده ورزی صورت پذیرفت. تحلیلگران با استفاده از این سیستم به تهیه گزارش‌ها در سناریوهای مختلف و ارزیابی افراد شرکت‌کننده به صورت فردی پرداختند. همه شرکت‌کنندگان در این ارزیابی متعلق به نیروی دریایی آمریکا بودند که از سه مرد و یک زن با درجه‌های مختلف استفاده شد. سناریوها شامل سؤالات پیچیده بود که چند سؤال شامل چندین زیر سؤال دیگر می‌شدند. ۲۱۲ گفتار توسط چهار تحلیلگر اطلاعاتی در طول مصاحبه فردی و دو گروه متمرکز به این دسته‌ها که کدگذاری شده بودند، ایجاد گردید. سپس با برنامه‌نویسی، دوازده دسته از مکالمات به هفت دسته کاهش پیدا کردند. در ادامه آن‌ها به افزونگی<sup>۱</sup>، نوآوری<sup>۲</sup>، تمامیت<sup>۳</sup>، اختصاصی بودن<sup>۴</sup>، گزینشی بودن<sup>۵</sup>، درک<sup>۶</sup>، ضریب اطمینان سیستم<sup>۷</sup>، انسجام<sup>۸</sup> و بافت<sup>۹</sup>، ادغام<sup>۱۰</sup> و کنترل<sup>۱۱</sup> روش خود پرداختند.

در مقاله دوم، هدف آن‌ها، توسعه روش‌های ارزیابی برای سیستم‌های IQA بود. در روش ارائه شده بحث در مورد طراحی و مدیریت ارزیابی قطعات و اثربخشی فن‌های ارزیابی گوناگون با توجه به اعتبار آن‌ها صورت پذیرفته بود. هدف از این کار فراهم کردن یک نقشه راه برای دیگران، تا بتوانند

---

<sup>1</sup> Redundency

<sup>2</sup> Novelty

<sup>3</sup> Completeness

<sup>4</sup> Specificity

<sup>5</sup> Selectiveness

<sup>6</sup> Comprehension

<sup>7</sup> Confidence in system

<sup>8</sup> Coherence

<sup>9</sup> Context

<sup>10</sup> Integration

<sup>11</sup> Control

سیستم‌های خود را ارزیابی و تحقیقات خود را به جلو پیش ببرند. در این مقاله کلی و همکارانش یک بررسی کلی بر روی یک سیستم تعاملی که آخرین تغییرات بین کاربر و سیستم را در نظر می‌گیرد (که بخشی از آن یا کلاً توسط کاربر کنترل می‌شود) انجام دادند. این مقاله حاصل از برگزاری دو هفته کارگاه بود. ارزیابی پیاده‌سازی شده توسط ۸ تحلیلگر، ۸ تا سناریو در حوزه سلاح‌های شیمیایی و کشتار جمعی، ۳ تا سیستم QA و یک سیستم بر پایه گوگل صورت پذیرفت. هر تحلیلگر از دو سیستم و ۲ سناریو برای آنالیز استفاده می‌کرد و در نهایت قرار بود که یک گزارش مربوط به قضاوت و ارزیابی خود ارائه نماید. البته این نکته قابل ذکر است که هر کدام از تحلیلگران به صورت جداگانه نحوه کار کردن با سیستم‌های QA را فراگرفته بودند. همچنین در نحوه انتخاب تحلیلگران سن، سطح تحصیلات، نوع شغل، تعداد سال خدمت در ارتش، تعداد سال انجام تجزیه و تحلیل، استفاده از کامپیوتر، تخصص در کامپیوتر و تجربه با سیستم پرس‌وجو مورد توجه قرار گرفته بود. دوره ارزیابی آن‌ها شامل چهار تا بلوک دو روزه، که هر بلوک به یک جفت تحلیلگر و به هر اتاق یک ناظر تخصیص داده شده بود، صورت پذیرفت. تحلیلگران با استفاده از دو دستگاه در هر اتاق به طور مستقل کار می‌کردند. سپس تحلیلگران و ناظران به صورت چرخشی در اتاق‌های مختلف با سیستم‌ها کار می‌کردند به طوری که تحلیلگران و ناظران تنها یک بار با هم بودند. در نهایت آن‌ها خلاصه‌ای از نکات کلیدی با توجه با اثربخشی این نکات، برای تخمین بین سیستم‌ها معرفی نمودند که از جمله می‌توان به موارد زیر اشاره کرد:

الف- مجموعه‌ای از نوشته‌ها<sup>۱</sup>: این مجموعه تخصصی و شامل CNS و اسناد وب است.

ب- سناریو<sup>۲</sup>: سناریوهای ارائه شده توسط مشاوران رومی AFRL تهیه شده بودند که برای تهیه یک سناریو نیاز به شناسایی محتوا و دامنه اطلاعات است.

ج- تحلیلگران<sup>۱</sup>: در اینجا تحلیلگران از نیروی دریایی انتخاب شده بودند.

<sup>۱</sup> Corpus  
<sup>۲</sup> Scenario

د-طراحی تجربی<sup>۲</sup>: هدف از این کار به دست آوردن نتایج آماری معتبرتر بود.

ه- مجموعه داده<sup>۳</sup>: فن‌های گسترده‌ای از جمع‌آوری داده در این مقاله استفاده شد.

و- فرضیه<sup>۴</sup>: در این مقاله فرض بر این بود که سیستم تعامل خوبی با کاربر دارد.

ز- تعامل<sup>۵</sup>

سان، روشی را برای ارزیابی سیستم‌های IQA معرفی نمود که X-EVAL نامیده می‌شد. آن‌ها سه سیستم توسعه یافته در ARDA<sup>۶</sup> برای برنامه AQUAINT<sup>۷</sup> را مورد مطالعه قرار دادند. آن‌ها در مقاله خود بیان داشتند که ارزیابی یک سیستم IQA به دلیل اینکه شامل دو موجودیت سیستم و فرد است، کار بسیار سخت و پیچیده‌ای می‌باشد. علت این امر ناشی از عوامل بسیاری است که یک کاربر تحت تأثیر آن‌ها با یک سیستم کار می‌نماید. بنابراین اگر چه روش‌های استاندارد وجود دارد که می‌توان اطلاعات مربوط به عملکرد سیستم از قبیل زمان، دقت و یا بازیابی را با استفاده از آن‌ها به دست آورد اما هنوز نیاز به شناسایی سهم سیستم و کاربران در عملکرد مطلوب یک سیستم است و تنها نتیجه‌گیری که می‌توانیم متصور شویم این است که عملکرد یک سیستم از کاربری به کاربر دیگر با ارزش‌تر خواهد بود. به عبارت دیگر روش ارزیابی باید قادر به بیان میزان سهم کاربر یا سیستم در موفقیت و یا شکست تولید خروجی مطلوب باشد. بنابراین آن‌ها در مقاله خود گزارش تجربی حاصل از مطالعه سه سیستم متفاوت IQA و یک سیستم پایه برای مقایسه ارائه کردند. ارزیابی‌ها بر روی همه سیستم‌ها توسط تحلیلگران انسانی صورت پذیرفت و از روش X-EVAL به عنوان روش سنجش اولیه مورد استفاده قرار گرفت. روش X-EVAL اولین بار در سال ۲۰۰۶ توسط سان و کانتور برای ارزیابی بازیابی اطلاعات مورد استفاده قرار گرفت. X-EVAL شاید اولین روش رسمی شد که به ارزیابی

<sup>1</sup> Analysts

<sup>2</sup> Experimental Design

<sup>3</sup> Data Collection

<sup>4</sup> Hypotheses

<sup>5</sup> Interaction

<sup>6</sup> Advance Research and Development Activity

<sup>7</sup> Advance Question Answering Intelligence

میزان مشارکت یک سیستم IQA برای رسیدن یک کاربر به نتایج نهایی پرداخت. قلب روش X-EVAL یک مدل خطی است که به طور خاص، جدا از اثرات اندازه‌گیری سیستم‌های تجربی، هر گونه اثر مختل کننده بالقوه، مانند توانایی تحلیلگر فردی، دشواری کار، اعمال نظر تحلیلگران به هنگام قضاوت محصولات کار و حتی امکان تعصب یک تحلیلگر در حمایت از کار خود و یا محصول او می‌باشد. معادله ۱-۲ این مدل را توصیف می‌نماید.

$$y(t, s, u, j, b) = \lambda^0 + \lambda_t^T + \lambda_s^S + \lambda_u^U + \lambda_j^J + \lambda_b^B + \varepsilon_{tsujb} \quad (1-2)$$

در حالی که  $y$  امتیاز محاسبه شده،  $T$  متغییر وظیفه،  $S$  متغیر سیستم،  $U$  متغیر کاربر،  $J$  متغیر قضاوت،  $B$  متغیر بایاس خود قضاوتی هست که اگر  $u=j$  باشد،  $b=I$  در غیر این صورت  $b=0$ ،  $\lambda_t^I$  تعیین سهم میزان متغیر مستقل  $I$ ،  $\lambda^0$  رهگیری (مقدار  $y$  زمانی که  $\lambda_t^I = 0$ ) و  $\varepsilon$  خطای تصادفی می‌باشند. این نکته قابل ذکر است که حروف کوچک ارائه شده در اینجا مقدار خاص از متغیر می‌باشند. آن‌ها تأکید داشتند که هدف از این گزارش ارزیابی سیستم نیست بلکه این کار را به منظور ارزیابی اثر X-EVAL در تشخیص پاسخ صحیح در میان این چهار سیستم بکار برده اند. روند ارزیابی در آزمایشگاه ملی Pacific Northwest وزارت دفاع آمریکا در شهر ریچلند انجام شده است. مدت زمان این کار در طول سه هفته بوده که یک هفته از این زمان برای آماده‌سازی سیستم‌ها و دو هفته برای ارزیابی روش ارائه شده زمان گذاشته شده است. همچنین برای روند انجام آزمایش یک کارگاه ارزیابی که شامل چهار بلوک دو روزه، هشت تحلیلگران اطلاعاتی و چهار اتاق بود که در هر بلوک، هر تحلیلگر به یکی از چهار سیستم تجربی اختصاص داده شده بود. به عبارت دیگر هر جفت از تحلیلگران اختصاص داده شده به سیستم QA در اتاقی کار می‌کردند که شامل دو ماشین با یک سیستم QA بود. از طرفی یک ناظر به هر جفت از تحلیلگران اختصاص داده شده بود. در طول فرآیند آزمایش صورت گرفته، تحلیلگران به طور مستقل بر روی دو سناریو در طول هر بلوک کار می‌کردند. روش ارائه شده بیشتر برای ارزیابی بازبایی اطلاعات بسیار حائز اهمیت بود [۴۵].

کوارترونی و ماناندهار روشی را که ارائه نمودند شامل یک ارزیابی کیفی از سیستم های پرسش و پاسخ تعاملی بود. آن ها در روش خود تعدادی پرسش مطرح کردند و از کاربران خواستند با دادن امتیازی بین یک (حداقل امتیاز) تا پنج (حداکثر امتیاز) کیفیت تعامل را اندازه گیری نمایند. سؤالاتی که آن ها در روش خود مطرح نمودند به شرح زیر بود.

- به چه میزان توانستند اطلاعات مورد نظر خود را از سیستم به دست آورید؟
- آیا سرعت سیستم در پاسخگویی به سؤالات شما مناسب بوده است؟
- آیا فکر می کنید سیستم در درک منظور شما موفق بوده است؟
- آیا کار با این سیستم به گونه ای بوده است که در آینده مجدداً از این سیستم استفاده نمایید؟
- آیا دسترسی به اطلاعات مورد نظرتان برای شما ساده بوده است؟
- آیا درخواست سیستم در بیان مجدد سؤال با کلمات یا ساختار جدید به نظرتان منطقی بوده است؟

• در کل، میزان رضایت خود از سیستم را اعلام نمایید؟

همان طور که از ظاهر سؤالات پیداست این پرسش ها شامل بررسی عملکرد سیستم، مشکلات تعامل، سرعت پاسخگویی و رضایت کلی کاربر از سیستم است که برای ارزیابی در نظر گرفته شده است [۴۶].

چیراگ شاه روشی را برای ارزیابی و پیش بینی کیفیت پاسخ در سیستم های پرسش و پاسخ اجتماعی<sup>۱</sup> ارائه نمودند که مورد توجه بسیاری از محققین این حوزه قرار گرفت. سیستم های QA توانسته اند به سیستم های پرس و جو مبتنی بر کلمات کلیدی و سیستم های بازیابی اطلاعات به شکل دقیق، از یک سند در مجموعه اسناد، کمک شایانی ذکر نمایند. بر این اساس چندین سیستم پرسش و پاسخ اجتماعی پدید آمد تا جستجوگران بتوانند توسط خدمات ارائه شده توسط آن ها، به اطلاعات مورد نیاز خود دسترسی پیدا کنند. در این سیستم ها ممکن است برای یک سؤال چندین پاسخ وجود داشته باشد و کسی که سؤال را مطرح کرده، بهترین پاسخ را از بین آن ها برای خود انتخاب نماید. در حالی که

<sup>۱</sup> Community Question Answering

نتایج نشان می‌داد که اکثر سؤال‌کننده‌ها، از دریافت پاسخ خود راضی هستند ولی هنوز روش روشنی برای ارزیابی کیفیت اطلاعات ارائه شده موجود نیست. کیفیت یک پاسخ یا هر اطلاعات شامل آن برای یک موضوع، می‌تواند ذهنی باشد اما تخصیص کیفیت می‌تواند وابسته به ارتباط با متن در مقابل دیگر فاکتورها باشد. بنابراین در این مقاله ارائه تفسیر در مورد کیفیت، بر اساس مجموعه داده موجود صورت پذیرفت. همچنین در این مقاله سعی بر پیش‌بینی این‌که آیا پاسخ انتخاب شده توسط کاربر می‌تواند به عنوان بهترین پاسخ باشد یا خیر بود. آن‌ها ابتدا یک مطالعه‌ای برای ارزیابی و پیش‌بینی کیفیت پاسخ در CQA ها انجام دادند سپس یاهو را به عنوان مرجع پاسخ‌ها و مجموعه کوچکی از سؤالات با پنج پاسخ آخر داده شده در مورد آن‌ها را انتخاب کردند. در ادامه از کارمندان شرکت آمازون خواستند که با توجه به ۱۳ شرط تعیین شده، به امتیازدهی هر کدام از این پاسخ‌ها بپردازند. برای هر پاسخ، آن‌ها از امتیاز پنج نفر استفاده کردند و در نهایت با ارزیابی امتیازات داده شده با امتیاز پرسش‌کننده جواب هر سؤال انتخاب شد. همچنین در این تحقیق آن‌ها با استخراج ویژگی‌های مختلف از تقویت پرسش، پاسخ، کاربران و آموزش تعداد طبقه بندی‌کننده برای انتخاب بهترین پاسخ نیز بهره گرفتند. آن‌ها در روش پیشنهادی خود، نشان دادند که اطلاعات مربوط به یک کاربر می‌تواند نقش مهمی در ارزیابی این سیستم‌ها داشته باشد. یاهو و AnswerBag از نمونه‌های سیستم‌های CQA هستند. در حوزه ادبیات مربوط به ارزیابی های CQA تعداد مقاله زیادی وجود ندارد اما بعضی از کارها وجود دارد که به پیش‌بینی میزان رضایت کاربران پرداخته است. اما کارهایی هم در زمینه بررسی عوامل افزایش کیفیت پاسخ خروجی سیستم نیز ارائه شده است. یکی دیگر از مزیت‌هایی که در ارزیابی این سیستم‌ها وجود داشت این بود که نه تنها سبب افزایش معیارهای انتخاب پاسخ مناسب برای کاربران می‌شود، بلکه سبب معروفیت سایت‌های سرویس دهنده در این زمینه می‌شود و این امر سبب می‌شود که با گرفتن رأی‌گیری در این سایت‌ها، مراجعه‌کنندگان بتوانند پاسخ‌های مناسب‌تری را دریافت نمایند. چیراگ در مقاله خود اظهار داشت که سیستم‌های CQA شامل سه مکانیسم ( الف) مکانیسم ثبت سؤال برای کاربران با زبان طبیعی ( ب) یک محل برای ارسال پاسخ به

کاربران ج) یک اجتماع یا جامعه ساخته شده برای این تغییرات) است. بسیاری از سایت‌های موجود در CQA مانند یاهو اجازه می‌دهند هر سؤالی، در هر زمینه‌ای، پرسیده شوند. ولی بعضی از سایت‌ها مانند Stack over flow موضوع محور می‌باشند و تنها به بحث در مورد موضوع خاصی می‌پردازند. در گزارش ارائه شده توسط چیراگ شاه و همکارانش دو تا روش برای تولید مجموعه ویژگی‌ها پیشنهاد شده است. یکی استفاده از یک مجموعه ۱۳ معیار برای شناسایی کیفیت و دیگری استفاده از ویژگی‌هایی که به صورت خودکار استخراج شده‌اند. برای تشکیل پایگاه داده از مجموعه پایگاه داده یاهو استفاده شده بود. در این مجموعه سؤالات به چهار گروه مشاوره، اطلاع‌رسانی، نظر و رأی‌گیری تقسیم و از این مجموعه به صورت تصادفی، ۳۰ سؤال از هر چهار دسته (برای هر سؤال پنج پاسخ آخر موجود بود) انتخاب گردید. سپس مجموعه داده ای با ۱۲۰ سؤال و ۶۰۰ پاسخ تشکیل شد. سپس با کمک ارزیاب‌های موجود در سیستم TREC از آن ارزیاب‌ها خواسته شد که این ۶۰۰ جواب را با توجه به ۱۳ معیار در نظر گرفته شده ارزیابی نمایند و یک امتیاز بین ۱ تا ۵ به این پاسخ‌ها تخصیص دهند. از آنجا که این معیارها ذهنی هستند، این احتمال وجود داشت که افراد مختلف، سطوح مختلف امتیاز را با توجه به شرایط در نظر بگیرند. بنابراین، از پنج ارزیاب مختلف برای به دست آوردن رتبه هر یک از این اظهارات استفاده شد که در کل ۳۰۰۰ ارزیابی مختلف ایجاد گردید. همان طور که در بالا تشریح شد هر ارزیاب قرار شد که بر اساس ۱۳ معیار بیان شده نمره‌ای بین ۱ تا ۵ به هر پاسخ تخصیص دهد. بنابراین برای تبدیل این امتیازات به دو کلاس بلی و خیر تصمیم گرفته شد که امتیازات بالاتر از ۳ جز کلاس بلی و امتیازات دیگر جز کلاس خیر قرار گیرند. بنابراین یک مدل رگرسیون منطقی برای پایگاه داده ایجاد گردید. در مرحله بعد آن‌ها از ۲۱ ویژگی استخراج شده به صورت خودکار برای پیش بینی مدل استفاده کردند و توانستند بر اساس این معیارها، دقتی برابر با ۸۴ درصد را گزارش نمایند. فرض مهمی که در ارائه بهترین پاسخ در نظر گرفته شد، عبارت بود از اینکه، پاسخی بهترین است برای یک سؤال اگر (۱) پاسخ توسط سؤال کننده به عنوان بهترین انتخاب شود و (۲)



پرسش‌کننده امتیاز آن را از ۳ یا بالاتر از آن در نظر بگیرد. این استاندارد طلایی بود که به آن‌ها کمک می‌کرد تا بتوانند یک مقایسه بین مدل برای ارزیابی و پیش‌بینی کیفیت پاسخ ارائه دهند [۴۷].

نینیا وچولدر و همکارانش در مقاله خود [۴۸] گزارشی از توسعه المان‌های یک طرح ارزیابی، برای طراحی کلی ارزیابی و قابلیت استفاده از سیستم HITIQA (که یک سیستم تعاملی پرسش و پاسخ برای تهیه گزارش گسترده در مسائل پیچیده است) ارائه کردند. در این گزارش دو هدف اساسی پیگیری شد. یکی انجام یک ارزیابی واقع‌بینانه از سودمندی و قابلیت استفاده از این سیستم به عنوان یک سیستم پایان به پایان<sup>۱</sup>، از سؤالات اولیه جستجوگر اطلاعات، برای تکمیل یک پیش‌نویس گزارش و دومین هدف توسعه معیارهای مقایسه پاسخ به دست آمده توسط تحلیلگران مختلف و ارزیابی کیفیت پشتیبانی HITIQA که آیا آن را فراهم می‌کند یا خیر. آن‌ها از ابزار کمی و کیفی برای به دست آوردن اطلاعات در مورد راحتی تحلیلگر با سیستم HITIQA استفاده کردند. به دلیل اینکه کیفیت خروجی اندازه‌گیری شده سیستم HITIQA با معیارهای استاندارد دقت و بازیابی سنجیده می‌شد. آن‌ها یک کار جدید (ارزیابی تقاطعی<sup>۲</sup>) برای اندازه‌گیری غیرمستقیم کیفیت پاسخ به دست آمده با استفاده از سیستم HITIQA را توسعه دادند و توانستند سیستم رأی دادن تحلیلگران به کیفیت خودشان و گزارش همکاران در ارزیابی سیستم را ارزیابی نمایند. آن‌ها برای ارزیابی، نتایج حاصل از دو سال توسعه این سیستم را در نظر گرفتند و از تحلیلگران اطلاعاتی دولت ایالات متحده برای شرکت در کارگاه‌های آموزشی دو سه روزه، دعوت به عمل آوردند. بنابراین آن‌ها در مقاله خود یک روش جدید برای ارزیابی از رأی تحلیلگران به کیفیت پاسخ فرد و گزارش همکاران به دست آمده بود، ارائه نمودند. آن‌ها همچنین ادعا نمودند که اگر سیستم از گزارش‌های تولید شده با کیفیت بالاتر استفاده کند می‌توان ادعا نمود که عملکرد سیستم بهتر شده است.

<sup>1</sup> End-to-End System

<sup>2</sup> Cross Evaluation

پاملا فورنر و همکارانش یک مرور کلی از مسائل مهم مطرح شده در طی هفت سال، بر روی یک سیستم دو زبانه QA انجام دادند. شرح مختصری از چگونگی انجام کار در طول این سال ها و نحوه تکامل مجموعه داده های ایجاد شده در مقاله آورده شده و در ادامه به ارائه خلاصه ای از سؤالات توسعه یافته مختلف پرداخته شده است. علاوه بر این، معیارهای ارزیابی، اصلی مورد استفاده برای ارزیابی عملکرد سیستم شرح داده شده و نتایج به دست آمده از تجزیه و تحلیل، ارائه شده است. آن ها در مقاله خود اظهار داشتند که در تحقیقات قبلی ارائه شده که بیشتر توسط کمپین TREC صورت پذیرفته است، ارزیابی توانایی سیستم ها در ارائه پاسخ دقیق در مورد سؤالات مطرح بوده و در این زمینه توانسته اند رشد قابل قبولی در ارزیابی سیستم های پرسش و پاسخ داشته باشند و توجه کمتری به ساختارهای سیستم های پرسش و پاسخ چند زبانی داشته اند. به عبارت دیگر، تحقیقات در زمینه سیستم های QA دو زبانه از مسیر اصلی محققین این حوزه خارج بوده است. اما در سال های اخیر توجه محققین به این حوزه به دو دلیل جلب شده است. دلیل اول به این خاطر است که کاربران به راحتی می توانند با سیستم های QA که به زبان اصلی فرد است ارتباط برقرار نموده و راحت تر کار نمایند و دوم اینکه دسترسی به اطلاعات تنها برای یک زبان خاص صورت می پذیرد. نکته قابل تأمل در این سیستم دو زبانه QA، این است که، در سیستم پیشنهادی QA برای CLEF، این سیستم با یک مجموعه سؤال مواجه بوده که از سیستم پرسیده شده و سیستم باید برای هر سؤال یک یا چند جواب دقیق را ارائه می کرد. آن ها برای ارزیابی سیستم خود، برای هر شرکت کننده در این آزمون باید برای هر سؤال حداقل یک جواب و یک سند پشتیبانی کننده از پاسخ ارائه می نمود. سپس هر پاسخ توسط یک انسان ارزیاب، ارزیابی می شد. قضاوت ارائه شده ممکن بود شامل یکی از پاسخ های زیر باشد.

- درست<sup>۱</sup> (R): رشته پاسخ شامل چیزی بیش از یک پاسخ دقیق نبود و آن حمایت می شد توسط یک متن پشتیبان.

- اشتباه<sup>۲</sup> (W): رشته پاسخ ارائه شده شامل جواب صحیح نیست.

- پشتیبانی نشده<sup>۳</sup> (U): پاسخ صحیح بود، اما برای پی بردن به صحت آن از متن پشتیبانی ارائه شده غیرممکن بود.

- نادرست<sup>۴</sup> (X): پاسخ صحیح و پشتیبانی می شد، اما رشته پاسخ موجود بیشتر یا کمتر از مقدار بیت پاسخ دقیق بود.

روش‌های ارزیابی ارائه شده در این مقاله شامل روش<sup>۴</sup> MRR<sup>۵</sup>، C@1، CWS<sup>۶</sup> و روش K1 بود [۴۹].

## ۲-۴ معیارهای ارزیابی موجود

بیشتر روش‌های ارزیابی موجود در زمینه سیستم‌های پرسش و پاسخ تعاملی، روش‌هایی هستند که در زمینه ارزیابی سیستم‌های پرسش و پاسخ و دیالوگ محور بکار گرفته شده‌اند [۵۰]. بیشتر این معیارها در مورد کیفیت پاسخ برگردانده شده و ارتقاء توانایی سیستم‌ها در برگرداندن پاسخ بوده و گزارشی از کیفیت سیستم‌ها ارائه نمی‌نمایند. در این بخش به معرفی تعدادی از این روش‌ها که برای ارزیابی سیستم‌های پرسش و پاسخ مورد استفاده قرار گرفته، پرداخته‌ایم. این روش‌ها به شرح زیر می‌باشند:

### ۲-۴-۱ روش MRR

زمانیکه سیستم برای پاسخ به سوال مطرح شده، چندین جواب را ارائه نماید از این معیار ارزیابی استفاده می‌گردد. بنابراین هنگامی که پرسش  $q_1$  تحویل مؤلفه بازبازی اطلاعات می‌شود، این مؤلفه

<sup>1</sup> Right

<sup>2</sup> Wrong

<sup>3</sup> Unsupported

<sup>4</sup> Inexact

<sup>5</sup> Mean Reciprocal Rank

<sup>6</sup> Confidence Weighted Score

زیرمجموعه‌ای از پاسخ‌ها را که به نام  $Sq_1$  شناخته می‌شود، بازیابی می‌کند. چنانچه در مجموعه

پاسخ‌های بازیابی شده، پاسخ صحیح، پاسخ  $j$ ام باشد، امتیاز پرسش با استفاده از فرمول  $Rank_{q_1} = \frac{1}{j}$

تعیین می‌شود. هنگامی که این محاسبه برای تمام پرسش‌ها در مجموعه  $Q$  انجام شود و میانگین

امتیازات کسب شده برای تک تک پرسش‌ها محاسبه گردد، مقدار جدیدی که به نام  $MRR$  شناخته

می‌شود، به دست می‌آید. این مقدار برابر با میانگین معکوس رتبه‌های پاسخ‌های ارائه شده صحیح

برای پرسش‌های مجموعه  $Q$  است.

$$MRR = \frac{1}{Q} \sum_{i=1}^Q Rank_{q_i} \quad (2-2)$$

عملکرد سیستم در ارائه پاسخ‌های صحیح با کمیت  $MRR$  ارتباط مستقیم دارد و با افزایش کمی

$MRR$ ، عملکرد سیستم بهتر خواهد بود، زیرا بالاتر بودن میزان  $MRR$  نشان دهنده بالاتر بودن امتیاز

محاسبه شده برای یکایک پرسش‌های  $Q$  و بنابراین کوچک تر بودن مخرج  $j$  در فرمول  $Rank_{q_1} = \frac{1}{j}$

است. کمتر بودن مخرج در این فرمول بدین معناست که پاسخ صحیح به ابتدای فهرست پاسخ‌های

بازیابی شده نزدیک تر بوده و بنابراین سرعت ارائه پاسخ صحیح به کاربر افزایش می‌یابد. یکی از

اشکالات این معیار این است که با استفاده از آن امکان ارزیابی پرسش‌هایی که همه پاسخ‌های

پیشنهاد شده برای آن‌ها نادرست بوده‌اند، وجود ندارد و بنابراین معیار  $MRR$  به میزان بازخوانی

سیستم توجه نمی‌کند. اشکال دیگر این است که نمی‌توان از این معیار در ارزیابی سیستم‌های پرسش

و پاسخی که برای هر پرسش فقط یک پاسخ ارائه می‌کنند استفاده نمود، زیرا در این گونه سیستم‌ها

فهرستی از پاسخ‌ها که بر اساس امتیاز مرتب شده‌اند، به کاربر ارائه نمی‌شود.

## ۲-۴-۲ روش C@1

در این روش همه سؤالات، تنها باید شامل یک پاسخ دقیق در مجموعه داده باشند. بنابراین سیستم

می‌تواند یک پاسخ به سؤال برگرداند و یا اینکه پاسخی را برای آن سؤال ارائه ننماید. در واقع در این

روش، یک نوع پاداش برای سیستم در نظر گرفته می شود تا ارائه پاسخ نادرست به سؤالات کم شود. این بدین معنا است، در صورتی که سیستم نتواند پاسخی برای سؤال ارائه نماید آن سؤال را رها کرده و این بهتر از این است که پاسخ اشتباهی را برای سؤال نشان دهد. در حقیقت یک استراتژی هست برای افزایش دقت سیستم. فرمول نحوه محاسبه این ارزیابی به شکل زیر است.

$$C @1 = \frac{1}{n} \left( n_R + n_u \frac{n_R}{n} \right) \quad (3-2)$$

در حالی که  $n_R$  برابر است با تعداد سؤالات صحیح پاسخ داده شده،  $n_U$  تعداد سؤالات پاسخ داده نشده و  $n$  تعداد کل سؤالات است.

### ۲-۴-۳ روش CWS

اولین بار این روش اندازه گیری توسط Voorhess [۵۱] معرفی شد. برای اندازه گیری این معیار، سیستم باید قادر به برگرداندن پاسخ ها بر اساس ضریب اطمینان باشد یعنی اینکه، سیستم بیشترین پاداش را به پاسخی می دهد که دارای بالاترین رتبه باشد. نحوه محاسبه این ارزیابی به شرح زیر است.

$$CWS = \frac{1}{n} \sum_{i=1}^n \frac{C(i)}{i} \quad (4-2)$$

$$C(i) = \sum_{j=1}^i I(j) \quad (5-2)$$

در حالی که  $n$  برابر با تعداد کل سؤالات،  $C(i)$  هست تعداد پاسخ های صحیح ارائه شده در موقعیت  $i$ ام در رتبه بندی و  $I(j)$  هست یک تابع که مقدار یک را بر می گرداند اگر جواب  $j$  صحیح باشد در غیر این صورت صفر را برمی گرداند.

### ۲-۴-۴ روش K1

این روش توسط Herrera در سال ۲۰۰۵ پیشنهاد گردید [۳]. سیستم هایی که بخواهند توسط این معیار مورد ارزیابی قرار گیرند باید یک عدد بین صفر و یک که نشان دهنده میزان ضریب اطمینان

آن‌ها به پاسخ است، ارائه نمایند. عدد یک بیانگر این است که سیستم کاملاً به پاسخ ارائه شده اطمینان دارد و صفر به معنی این است که سیستم به صحت پاسخ ارائه شده اطمینان ندارد. K1 بر مبنای یک سیستم نرم‌افزاری عمل می‌نماید. در این سیستم نرم‌افزاری، در صورتی که پاسخ نادرست باشد عدد ۱- را بر گردانده می‌شود و در صورتی که پاسخ درست باشد عدد ۱ را برمی‌گرداند. در نهایت این اعداد بر گردانده شده با ضرایب اطمینان برگردانده شده توسط سیستم وزن دهی می‌شوند و یک عددی بین ۱- تا ۱ به عنوان خروجی ارائه می‌گردد. تفسیر این معیار مقداری مشکل است. بدین معنا که یک مقدار مثبت لزوماً به معنای ارائه پاسخ‌های صحیح بیش از آن‌هایی که اشتباه نشان نمی‌دهد، نیست، اما در مجموع نمرات پاسخ‌های صحیح، بالاتر از مجموع نمرات آن‌هایی که اشتباه شده‌اند، می‌باشد. فرمول زیر نحوه محاسبه K1 را نمایش می‌دهد.

$$K1 = \frac{\sum_{i \in (\text{correct\_answer})} \text{self\_score}(i) - \sum_{i \in (\text{incorrect\_answer})} \text{self\_score}(i)}{n} \quad (6-2)$$

## ۲-۴-۵ بازخوانی<sup>۱</sup> و دقت<sup>۲</sup>

دو معیار بازخوانی و دقت، جهت ارزیابی کارایی<sup>۳</sup> سیستم مورد استفاده قرار می‌گیرند. بازخوانی یعنی از میان کل پاسخ‌های صحیح موجود در مجموعه اسناد، چند درصد از آن‌ها یافت شده‌اند و دقت یعنی در میان پاسخ‌های یافت شده، چند درصد صحیح می‌باشند. دقت و بازخوانی با استفاده از روابط زیر قابل اندازه‌گیری هستند.

$$\text{Recall} = \frac{r}{R} \quad (7-2)$$

$$\text{Precision} = \frac{r}{N} \quad (8-2)$$

<sup>1</sup> Recall

<sup>2</sup> Precision

<sup>3</sup> Performance

در حالی که  $r$  عبارتست از تعداد مربوط به حقایق بازیابی شده،  $R$  تعداد کل حقایق موجود در کلید پاسخ و  $N$  مجموع تعداد حقایق در پاسخ سیستم می باشد. همچنین تعریف دیگری از دقت و بازخوانی به شکل زیر است.

$$Recall = \frac{TP}{TP + FN} \quad (۹-۲)$$

$$Precision = \frac{TP}{TP + FP} \quad (۱۰-۲)$$

که در آن  $TP$  بازیابی شده مرتبط،  $FN$  بازیابی نشده مرتبط و  $FP$  بازیابی شده نامرتب است. مقدار این دو پارامتر رابطه معکوسی با هم دارند و افزایش مقدار یکی باعث کاهش و افت دیگری می شود. از این رو بایستی مصالحه‌ای بین این دو مقدار انجام شود. بین صحت و توانایی بازخوانی سیستم رابطه معکوس وجود دارد و این درحالیست که برای بسیاری از کاربران صحت سیستم ممکن است نسبت به توانایی بازخوانی آن، به میزان  $B$  برابر ارزش بیشتری داشته باشد. با توجه به این مسئله چنانچه برای معیار صحت وزن یا ارزش  $B$  قائل شویم، معیار اف که ترکیبی از معیارهای صحت و بازخوانی و در حقیقت میانگین هماهنگ دو پارامتر فوق است، به صورت زیر محاسبه می شود.

$$F(\beta) = \frac{(\beta^2 + 1) * Precision * Recall}{\beta^2 * Precision + Recall} \quad (۱۱-۲)$$

با استفاده از این معیار می توان صحت و بازخوانی را در مقابل هم تعدیل کرد. این معیار به ویژه زمانی که پاسخ‌های مرتبی در نظام بازیابی اطلاعات هستند، معمولاً به معیار صحت ترجیح داده می شود و از معیارهای معمول ارزیابی بخصوص در مواقع کار با مجموعه‌های غیرمتعادل است. تعیین حداکثر مقدار برای  $F$  به منظور یافتن بهترین ترکیب ممکن بین جامعیت و مانعیت است.

## ۲-۴-۶ دقت در $n$ نتیجه اول ( $p@n$ )

در تعیین معیار دقت، تمامی پاسخ‌های بازگردانده شده توسط سیستم مورد توجه قرار می گیرند و بنابراین این معیار اطلاعات مناسبی درباره تعداد پاسخ‌های صحیح موجود در  $n$  پاسخ ابتدای فهرست

پیشنهادی پاسخ‌ها ارائه نمی‌کند. به عنوان مثال، ممکن است در فهرست پاسخ‌های پیشنهادی توسط یک سیستم پاسخ‌های صحیح در رتبه‌های بالاتر و در ابتدای فهرست ارائه شده باشند و در سیستم دیگر همین تعداد پاسخ صحیح در رتبه‌های پایین و در اواسط یا اواخر فهرست پاسخ‌های صحیح قرار داشته باشند. در این حالت معیار دقت برای دو سیستم ارزش یکسانی در نظر می‌گیرد. برای تشخیص تفاوت این دو سیستم از نظر دقت پاسخ‌های ارائه شده باید از معیار دیگری که  $p@n$  نامیده می‌شود، استفاده نمود. برای ارزیابی توانایی بازخوانی این دو سیستم نیز تنها  $n$  نتیجه ابتدای فهرست بازبایی شده که دارای بیشترین امتیاز هستند، در نظر گرفته می‌شوند.

$$precision^s_{(Q,D,n)} = \frac{\sum_{q \in Q} \frac{|R^s_{(q,D,n)} \cap A_{(q,D)}|}{|R^s_{(q,D,n)}|}}{|Q|} \quad (12-2)$$

$$Recall^s_{(Q,D,n)} = \frac{\sum_{q \in Q} \frac{|R^s_{(q,D,n)} \cap A_{(q,D)}|}{|A_{(q,D)}|}}{|Q|} \quad (13-2)$$

## ۲-۴-۷ افزونگی<sup>۱</sup>

این معیار نشان دهنده میانگین تعداد پاسخ صحیح در مجموعه پاسخ‌های بازگردانده شده برای هر پرسش است. به بیان دیگر این معیار احتمال اینکه مؤلفه استخراج اطلاعات بتواند برای پرسش مطرح شده، حداقل یک پاسخ صحیح بیابد را محاسبه می‌کند.

$$Redundancy^s_{(Q,D,n)} = \sum_{q \in Q} \frac{|R^s_{(q,D,n)} \cap A_{(q,D)}|}{|Q|} \quad (14-2)$$

این نکته قابل ذکر است که در روابط بالا متغیرهای بکار گرفته شده به شرح زیر می‌باشند:

$Q$ : به مجموعه پرسش‌هایی که به مؤلفه بازبایی اطلاعات در سیستم پرسش و پاسخ داده می‌شود اشاره می‌کند.

$q$ : نشان دهنده هر یک از پرسش‌های موجود در مجموعه پرسش‌ها است.

<sup>۱</sup> Redundancy



$D$ : نشان دهنده مجموعه متون موجود در منابع اطلاعات سیستم است.

$A(q, D)$ : به مجموعه قطعات متنی در منبع اطلاعات که پاسخ‌های صحیحی برای پرسش  $q$  می‌باشند اشاره می‌کند.

$S$ : زیرمجموعه‌ای از اسناد موجود در منبع اطلاعاتی  $D$  است که در ارتباط با پرسش کاربر ارزیابی می‌شوند.

$R_{(q,D,n)}^s$ : هنگامی که پرسش  $q$  به مؤلفه بازیابی اطلاعات وارد می‌شود، مجموعه اسناد  $S$  از منبع اطلاعاتی  $D$  بازیابی می‌شوند. در این مجموعه،  $n$  سند یا قطعه متن بازیابی شده که دارای بیشترین امتیاز هستند، نامیده می‌شوند.

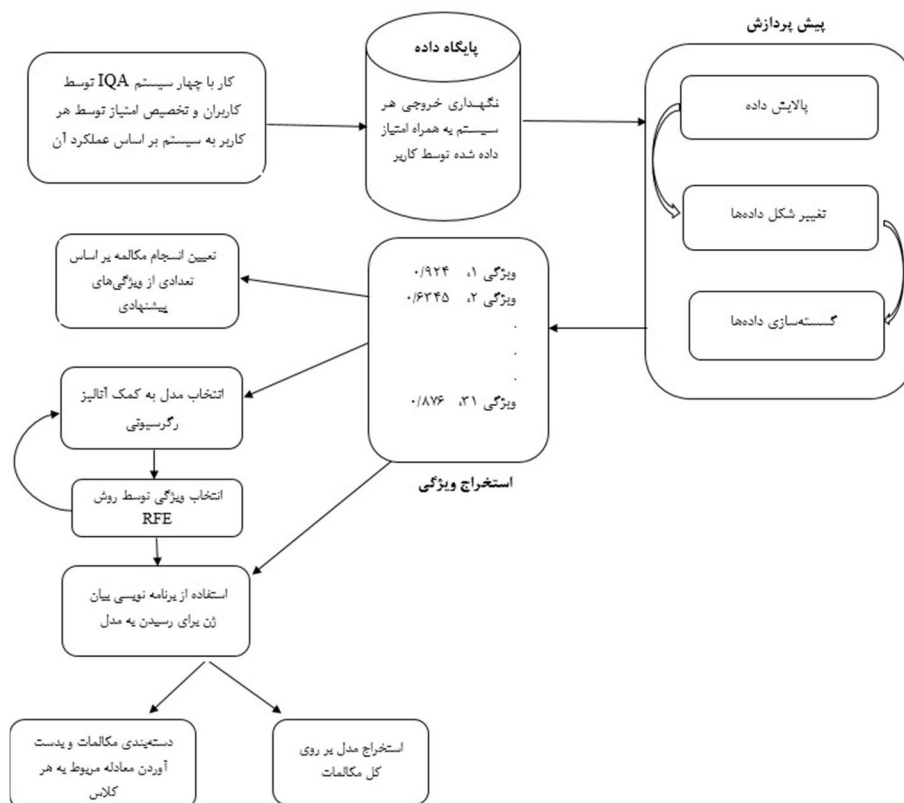


## فصل سوم-روش پیشنهادی

در یک سیستم QA، کارایی یک سیستم بر اساس چگونگی برگرداندن یک پاسخ مناسب، در برابر یک پرسش مطرح شده از یک حوزه خاص، مورد ارزیابی قرار می‌گیرد. در حالی که در سیستم‌های پرسش و پاسخ تعاملی، بدلیل اینکه کاربران برای رسیدن به اطلاعات مورد نیاز خود با سیستم در تعامل هستند، کارایی سیستم‌های IQA به طور سنتی به وسیله آنالیز جنبه‌های مختلف آن‌ها مورد ارزیابی قرار می‌گیرد. ارزیابی سیستم‌های IQA به منظور تعیین و ارتقای کارایی آن‌ها از اهمیت زیادی برخوردار است. با این وجود بررسی تحقیقات گذشته نشان داده هنوز روش استاندارد و خاصی برای ارزیابی این سیستم‌ها ارائه نشده است و ارزیابی‌ها معمولاً با استفاده از روش‌های مرسوم در ارزیابی سیستم‌های QA و نیز سیستم‌های دیالوگ محور صورت می‌گیرد [۴]. در ارزیابی سیستم‌های IQA، علاوه بر ارزیابی کمی از ارزیابی کیفی نیز استفاده می‌شود که این امر نیازمند مشارکت کاربران در فرآیند ارزیابی، برای تعیین میزان موفقیت تعامل بین سیستم و کاربر است. به دلیل مشارکت افراد در فرآیند ارزیابی، ارزش یک سیستم از کاربری به کاربر دیگر متفاوت خواهد شد لذا برای جلوگیری از این اتفاق در ارزیابی سیستم‌های IQA از چند ارزیاب استفاده می‌شود [۵۲]. بنابراین اگر بتوان مدلی پیشنهاد نمود که بتواند جایگزین ارزیاب‌های انسانی شود می‌توان گام مهمی در فرآیند توسعه و ارزیابی این سیستم‌ها برداشت. لذا در این فصل به معرفی یک مدل آماری جهت ارزیابی سیستم‌های پرسش و پاسخ تعاملی خواهیم پرداخت.

### ۳-۱ روش و طرح کلی تحقیق

هدف از این تحقیق، طراحی یک روش مستقل از زبان و حوزه معنایی به منظور ارزیابی یک سیستم پرسش و پاسخ تعاملی است بطوریکه، مدل پیشنهادی بتواند جایگزین ارزیاب‌های انسانی شود. مدل پیشنهادی بدین صورت در نظر گرفته شده که خروجی آن بیشترین شباهت را با نظرات ارزیاب‌ها داشته باشد به عبارتی مدل طوری طراحی شده که کمترین خطا را نسبت به نظرات موجود داشته باشد. شکل ۳-۱ فرآیند رسیدن به مدل پیشنهادی را نمایش می‌دهد.



شکل ۳-۱ فرآیند شکل‌گیری روش ارزیابی پیشنهادی

در ارائه یک روش ارزیابی کامل، نیاز به شناسایی تمامی ابعاد یک مدل ارزیابی است. از آنجایی که یک سیستم پرسش و پاسخ تعاملی از دو موجودیت سیستم و کاربر تشکیل شده و تعدد ویژگی‌ها از جنبه‌های مختلف، امکان ایجاد یک روش استاندارد و جامع را بسیار سخت می‌سازد. بنابراین امکان ارائه یک روش کاملاً خودکار به منظور ارزیابی این سیستم‌ها امکان‌پذیر نبود ولی می‌توان با خودکار کردن بسیاری از ویژگی‌ها و جایگزین کردن یک مدل به جای انسان تا حدود زیادی به فرآیند ارزیابی این سیستم‌ها کمک نمود. تکنیک‌های آماری استفاده شده، علاوه بر ایجاد یک سیستم ارزیابی مستقل از زبان، سرعت فرآیند ارزیابی را نیز افزایش می‌دهد. با این وجود کاهش دقت ارزیابی با توجه به عدم استفاده از دانش زبان شناختی و ارزیاب انسانی امری محتمل است. به منظور نیل به اهداف فوق، ابتدا چهار سیستم پرسش و پاسخ تعاملی موجود در نظر گرفته شد. سپس پایگاه داده‌ای از مکالمات رد و بدل شده بین کاربران و سیستم‌ها تشکیل گردید. بعد از انجام مرحله پیش پردازش متون، استخراج

ویژگی‌های آماری صورت پذیرفت و در نهایت مدل رگرسیونی با توجه به ویژگی‌های مطرح شده، حاصل گردید که در ادامه توضیح هر یک از این مراحل به تفصیل آورده شده است.

### ۳-۲ پیش پردازش

یکی از مراحل اولیه برای متن کاوی پیش پردازش است. پیش پردازش، اولین گام در جهت تطابق مستندات متنی با نمایش آنها در یک قالب مناسب می‌باشد. در حقیقت، پیش پردازش وظیفه نگاشت متن داده شده به یک نمای منطقی را بر عهده دارد. پردازش زبان فارسی از جهاتی با پردازش زبان انگلیسی تفاوت دارد. در زبان انگلیسی تمامی حروف و تمامی کلمات جدا از هم و با قانونی مشخص نوشته می‌شوند. در حالی که در زبان فارسی بعضی از حروف به هم چسبیده یا جدا از هم نوشته می‌شوند، بعضی از کلمات یکپارچه، با فاصله یا نیم فاصله به دو یا چند بخش تقسیم می‌شوند. شکل‌های غیراستاندارد نویسه‌ها و کلمات به وفور در متون نوشته دیده می‌شوند. لذا قبل از این که بتوان از این متون استفاده کرد و یا در پایگاه داده ذخیره نمود، باید ابتدا پیش پردازشی روی آنها صورت پذیرد تا شکل غیراستاندارد به شکل استاندارد تبدیل گردد. در این مسیر، بایستی از یک استاندارد مشترک برای پیش پردازش و پردازش متون استفاده کرد. پیش‌پردازش اطلاعات در ۵ مرحله صورت پذیرفت.

۱- مشخص کردن مرز جمله‌ها: در بیشتر مواقع، تعیین مرز جمله‌ها از طریق بررسی علائم جداکننده از قبیل فضای خالی، "، "!"، "؟"، "؛" و غیره انجام می‌شود. همچنین برای متون انگلیسی علاوه بر این علائم از تجزیه‌کننده استنفورد<sup>۱</sup> استفاده شد.

۲- ریشه‌یابی: در این حالت یک کلمه به شکل عمومی خود کاهش می‌یابد که این شکل عمومی باید برای همه کلمات هم‌ریشه یکسان باشد. برای دادگان انگلیسی از ریشه‌یاب استنفورد و مجموعه دادگان فارسی از ابزارهای پردازش متون دانشگاه فردوسی<sup>۲</sup> استفاده گردید.

<sup>۱</sup> <https://nlp.stanford.edu/software/lex-parser.shtml>

<sup>۲</sup> [www.wtlab.um.ac.ir](http://www.wtlab.um.ac.ir)

- ۳- حذف کلمات و واژه‌های غیرمهم: در این مرحله لیستی مشتمل بر ۲۰۰ کلمه پرتکرار (ایست واژه<sup>۱</sup>) آماده گردید (کلماتی که در محتوای اصلی متن تاثیری ندارند) و از مکالمه‌ها حذف گردید.
- ۴- شناسایی مقادیر عددی: بعد از شناسایی اعدادی که به صورت حروف در مکالمه‌ها ذکر شده بودند، این کلمات برچسب مقدار عددی دریافت کردند.
- ۵- یکسان‌سازی متن‌ها: در متون انگلیسی تمامی کلمات با حروف بزرگ به حروف کوچک تبدیل شدند و در متون فارسی یکسان‌سازی حروف (مثل حروف "ی" و "ک") صورت پذیرفت.

### ۳-۳ استخراج ویژگی

یکی از مهم‌ترین قسمت‌های مربوط به هر سیستم تشخیص یا مدل‌سازی، استخراج ویژگی است. با انجام دقیق این مرحله، نتایج حاصل از مراحل بعدی دارای دقت بالاتری خواهد بود. اولین گام برای ارائه یک مدل آماری شناسایی ویژگی‌های لازم در ارزیابی یک سیستم پرسش و پاسخ تعاملی و سپس به دست آوردن ضرایب هر یک از پارامترهای ارزیابی است. بنابراین در ابتدا می‌بایست ویژگی‌های شاخص را فهرست نموده سپس به بررسی و محاسبه تأثیر هر یک از آنها در ارزیابی سیستم‌های پرسش و پاسخ تعاملی بپردازیم. با توجه به مطالعات صورت گرفته در حوزه ارزیابی، معمولاً پارامترهای متفاوتی از دیدگاه ارزیاب‌ها مورد توجه قرار می‌گیرد [۵۳].

جدول ۳-۱ فهرست جمع‌آوری شده از ویژگی‌هایی که در ارزیابی سیستم‌های پرسش و پاسخ تعاملی که معمولاً در تحقیقات قبلی مورد استفاده قرار گرفته‌اند، را نشان می‌دهد. این ویژگی‌ها تمامی ابعاد مختلف ارزیابی را شامل می‌شوند. اما همان‌طور که قبلاً بیان شد ارائه یک مدل کلی که به ارزیابی خودکار سیستم‌های تعاملی بپردازد کار بسیار سختی است. لذا ما تنها به خودکار کردن ویژگی‌هایی پرداختیم که توسط ارزیاب انسانی در امتیازدهی به خروجی سیستم‌ها در نظر گرفته می‌شود.

<sup>1</sup> Stop-words

از ویژگی‌های مهمی که توسط ارزیاب‌ها در نظر گرفته می‌شود می‌توان به تطبیق نام موجودیت<sup>۱</sup> (ویژگی که مشخص می‌کند که آیا تمام موجودیت‌های ظاهر شده به طور مثال در پرسش  $Q_{i+1}$  در پرسش  $Q_i$  قرار دارند یا نه)، تطبیق هدف<sup>۲</sup> (تعیین کننده میزان شباهت به طور مثال بین نوع پرسش  $Q_{i+1}$  و پرسش  $Q_i$  است)، فهم پرسش، پوشش<sup>۳</sup>، گستردگی<sup>۴</sup>، تمامیت<sup>۵</sup>، ارتباط<sup>۶</sup>، دقت<sup>۷</sup> و حجم<sup>۸</sup> اشاره نمود. هر کدام از روابط معرفی شده در این قسمت سعی بر پوشش این ابعاد از نظرات ارزیاب‌ها را دارند. برخی از ویژگی‌های پیشنهادی تنها یک بعد و برخی از ویژگی‌ها چندین بعد از دیدگاه ارزیاب‌ها را پوشش می‌دهد. در این مرحله، ویژگی‌های حاصل گردید که در ادامه به معرفی هر یک از آن‌ها خواهیم پرداخت. با توجه به اینکه خروجی هر مکالمه صورت گرفته بین کاربران و سیستم‌ها به صورت مجموعه‌ای از سؤال‌ها و پاسخ‌ها آماده گردید، برخی از ویژگی‌های تعریف شده علاوه بر اعمال بر یک جفت پرسش-پاسخ، به صورت جداگانه برای مجموعه پرسش و همینطور مجموعه پاسخ‌ها نیز به صورت جداگانه بکار گرفته شدند. یعنی مکالمات را به صورت مجموعه‌های مختلف در نظر گرفته و هر یک از ویژگی‌های حاصل را بر روی آن‌ها محاسبه کردیم. این کار به این دلیل انجام می‌شود که معمولاً یک مکالمه دنباله‌ای از پرسش و پاسخ‌ها است که به یکدیگر مرتبط هستند. به عبارتی دیگر مکالمه‌ای دارای انسجام بیشتر است که سؤالات و پاسخ‌های رد و بدل شده ارتباط بیشتری با یکدیگر داشته باشند. با توجه به اینکه امکان پیاده‌سازی به صورت خودکار همگی این ویژگی‌ها وجود نداشت. لذا سعی گردید ویژگی‌های پیشنهادی به صورتی باشد که بتواند بخشی از این ویژگی‌ها را پوشش دهد. ویژگی‌های پیشنهاد شده در ادامه آورده شده است:

---

<sup>1</sup> Named entity matching

<sup>2</sup> Target Matching

<sup>3</sup> Coverage

<sup>4</sup> Extensiveness

<sup>5</sup> Completeness

<sup>6</sup> Relevance

<sup>7</sup> Accuracy

<sup>8</sup> Size



جدول ۳-۱ لیست ویژگی‌های در نظر گرفته شده جهت ارزیابی سیستم‌های پرسش و پاسخ

ردیف	دسته ویژگی	ویژگی‌ها
۱	کارایی سیستم	سرعت <sup>۱</sup> ، در دسترس بودن <sup>۲</sup> ، قابلیت اطمینان <sup>۳</sup>
۲	پاسخ‌ها	تمامیت <sup>۴</sup> ، ارتباط <sup>۵</sup> ، دقت <sup>۶</sup> ، سودمندی <sup>۷</sup> ، قابلیت اجرا <sup>۸</sup>
۳	محتوی پایگاه داده	صلاحیت <sup>۹</sup> ، منشأ <sup>۱۰</sup> ، کیفیت منبع <sup>۱۱</sup> ، محدوده <sup>۱۲</sup> ، گستردگی <sup>۱۳</sup> ، پوشش <sup>۱۴</sup> ، حجم <sup>۱۵</sup> ، به روز رسانی <sup>۱۶</sup>
۴	نحوه نمایش (اعم از ورودی و خروجی)	فهم سؤال، اطلاعات مورد نیاز فهم، سبک جستجو، سؤال، کلمات کلیدی، مخفف و ....
۵	مطابقت با هدف <sup>۱۷</sup>	تعیین کننده میزان شباهت بین نوع پرسش $Q_{i+1}$ و پرسش $Q_i$ است.
۶	تطبیق نام موجودیت <sup>۱۸</sup>	ویژگی که مشخص می‌کند که آیا تمام موجودیت‌های ظاهر شده در پرسش $Q_{i+1}$ در پرسش $Q_i$ قرار دارند یا نه
۷	میزان شباهت بین پرسش‌ها	ویژگی است که میزان شباهت بین پرسش $Q_{i+1}$ و پرسش $Q_i$ را نشان می‌دهد.
۸	شباهت بین متن کلمات سؤالات	این ویژگی میزان شباهت بین متن کلمات سؤالات را اندازه‌گیری می‌نماید.
۹	شباهت بین پرسش و پاسخ	ویژگی که بیانگر شباهت بین پرسش $Q_i$ و پاسخ $A_i$ است.
۱۰	شباهت بین متن پرسش و پاسخ	ویژگی که بیانگر شباهت بین پرسش $Q_i$ و پاسخ $A_i$ است بر اساس متن کلمات (به استثنای نام موجودیت‌ها)
۱۱	دیگر ویژگی‌ها	طول موضوع سؤال، طول محتوای سؤال، تعداد پاسخ به این پرسش، میزان شباهت دو پاسخ داده شده پشت سر هم، تعداد سؤالات و پاسخ‌های رد و بدل شده برای رسیدن به جواب، زمان پاسخگویی و ....

<sup>1</sup> Speed

<sup>2</sup> Availability

<sup>3</sup> Reliability

<sup>4</sup> Completeness

<sup>5</sup> Relevance

<sup>6</sup> Accuracy

<sup>7</sup> Usefulness

<sup>8</sup> Applicability to task

<sup>9</sup> Authority

<sup>10</sup> Provenance

<sup>11</sup> Source quality

<sup>12</sup> Scope

<sup>13</sup> Extensiveness

<sup>14</sup> Coverage

<sup>15</sup> Size

<sup>16</sup> Updatedness

<sup>17</sup> Target Matching

<sup>18</sup> Named entity matching

## - ویژگی اول:

N-gram ها یکی از مشهورترین مدل‌های آماری زبان هستند. در این مدل‌ها ارتباطات زنجیره کلمات در نظر گرفته می‌شود. به عبارت دیگر، مدل‌های N-gram بر اساس هم پیوندی و کنار هم قرار گرفتن کاراکترهای لغات در پردازش متن عمل می‌نمایند. ابتدا N-gram های مشترک را شمرده با یکدیگر جمع و بر تعداد کل N-gram ها تقسیم می‌نماییم (رابطه ۱-۳).

$$X_1 = \sum_{S_i \in conv} \frac{\sum_{ngram \in S_i} Count_{match}(gram_n)}{\sum_{ngram \in S_i} Count(gram_n)} \quad (1-3)$$

که در این رابطه  $S_i$ ،  $i$ -امین جمله از هر مجموعه مکالمه (conv) و  $n$  طول هر N-gram است. فرض نمایید یک مکالمه شامل  $N$  پرسش و پاسخ باشد. N-gram های مشترک بین هر پرسش-پاسخ را شمرده با یکدیگر جمع و بر مجموع تعداد N-gram ها تقسیم می‌نماییم. این کار به ازای  $n=1,2,3$  برای مجموعه‌های پرسش و پاسخ  $(Q_i-A_i)$ ، مجموعه پرسش‌ها  $(Q_i-Q_{i+1})$  و مجموعه پاسخ‌ها  $(A_i-A_{i+1})$  از هر مکالمه صورت پذیرفت.

## - ویژگی دوم:

در یک مکالمه برای  $n$  های بزرگ‌تر، هر چه تعداد N-gram های مشترک افزایش یابد، امتیاز آن مکالمه بیشتر و احتمال پیوستگی متن مکالمه بیشتر خواهد شد [۵۳]. بر این اساس در ویژگی پیشنهادی، هر کدام از N-gram ها، بر اساس یک ضریب وزنی برای هر N-gram به ارزش  $W_i$  با یکدیگر جمع می‌شوند تا مقدار این ویژگی به دست آید (رابطه ۲-۳).

$$X_2 = \frac{1}{M} \times \sum_{i=1}^M \frac{\sum_{ngram \in S_i} W_n \times Count_{match}(gram_n)}{\sum_{ngram \in S_i} Count(gram_n)} \quad (2-3)$$

که در آن  $M$  تعداد عضوهای مجموعه جهت محاسبه میانگین و  $W_n$  ضریب تأثیر هر N-gram و مقدار آن متناسب با عدد  $n$  است. این ویژگی نیز برای  $n=1,2,3$  محاسبه شده است. نحوه محاسبه این رابطه مانند ویژگی اول است با این تفاوت که به ازای N-gram های مشترک ضریبی برابر با  $n$  به هر N-gram مشترک نسبت داده می‌شود.

### - ویژگی سوم:

متناسب با رابطه ۲-۳، رابطه ۳-۳ پیشنهاد گردید، در این ویژگی، ابتدا به ازای هر جفت پرسش-پاسخ، ابتدا تعداد N-gram های مشترک برای یک زوج از هر مجموعه (مثلاً یک پرسش و پاسخ) را به ازای  $n$  های مختلف محاسبه و متناسب با مقدار  $n$ ، تعداد را در ضریب ارزش هر N-gram ضرب، سپس بر مجموع تعداد N-gram ها تقسیم نموده، ماکزیمم بین آن‌ها را در نظر گرفته شد. با توجه به این که هر مکالمه از  $M$  تا زوج تشکیل شده است. در نهایت پاسخ به دست آمده از میانگین‌گیری بین  $M$  تا زوج درون یک مکالمه حاصل گردید.

$$X_3 = \frac{1}{M} \times \sum_{i=1}^M \arg \max_{ngram \in S_i} \left( \frac{W_n \times Count_{match}(gram_n)}{\sum_{ngram \in S_i} Count(gram_n)} \right) \quad (3-3)$$

### - ویژگی چهارم:

انطباق پشت سر هم در سطح جمله معمولاً در N-gram ها دیده می‌شود. بنابراین در N-gram ها طول تعریف نمی‌شود. زیرا بزرگ‌ترین رشته مشترک در نظر گرفته می‌شود. در حالیکه در بزرگ‌ترین زیر رشته مشترک نیاز نیست انطباق پشت سر هم باشد. همچنین برای اینکه مسئله هم‌رخدادی در جملات نیز در نظر گرفته شود، از معادله ۳-۶ استفاده نمودیم. در رابطه تعریف شده، برای یک مکالمه ابتدا یک زوج پرسش-پاسخ را در نظر گرفته، سپس برای هر جفت بازایی و دقت را محاسبه و برای تمامی زوج پرسش-پاسخ این کار را انجام می‌دهیم. در نهایت پاسخ به دست آمده را در رابطه ۳-۶ قرار داده و امتیاز هر مکالمه را محاسبه می‌کنیم.

$$R_{LCS} = \frac{1}{M} \times \sum_{i=1}^M \frac{LCS(Q_i, A_i)}{L_{Q_i}} \quad (4-3)$$

$$P_{LCS} = \frac{1}{M} \times \sum_{i=1}^M \frac{LCS(Q_i, A_i)}{L_{A_i}} \quad (5-3)$$

$$X_4 = \frac{(1 + \beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (6-3)$$

که در آن  $\beta = \frac{P_{LCS}}{R_{LCS}}$ ،  $M$  تعداد جفت سوال-پاسخ هر مکالمه،  $LCS(Q_i, A_i)$  بزرگ‌ترین زیر رشته

مشترک بین سؤالات و پاسخ‌های یک مکالمه و  $L$  طول سؤال یا جواب است.

#### - ویژگی پنجم:

در این ویژگی اجتماع بزرگ‌ترین زیر رشته مشترک بین  $Q_i$  و مجموعه جواب‌ها را محاسبه گردید. عدد بزرگ‌تر بدست‌آمده از این ویژگی بیانگر ارتباط بیشتر بین جملات مکالمه خواهد بود. به‌طور مثال فرض کنید جمله  $Q_1$  شامل کلمات  $w_1 w_2 w_3 w_4 w_5$  و پاسخ  $A_1$  شامل کلمات  $w_1 w_2 w_6 w_7 w_8$  و پاسخ  $A_2$  شامل کلمات  $w_1 w_3 w_8 w_9 w_5$  باشد. LCS در رابطه بین  $Q_1$  و  $A_1$  برابر  $w_1 w_2$  و بین  $Q_1$  و  $A_2$  برابر  $w_1 w_3 w_5$  است. لذا اجتماع بین  $Q_1$ ،  $A_1$  و  $A_2$  برابر با  $w_1 w_2 w_3 w_5$  است که  $LCS_{\cup}(Q, A) = \frac{4}{5}$  حاصل می‌گردد. برای کل مجموعه سؤالات یک مکالمه این کار را انجام دادیم. بنابراین روابط به صورت زیر پیشنهاد گردید.

$$R_{LCS} = \frac{1}{M} \times \sum_{i=1}^M \frac{\sum_{j=1}^U LCS_{\cup}(Q_i, A_j)}{P} \quad (7-3)$$

$$P_{LCS} = \frac{1}{M} \times \sum_{i=1}^M \frac{\sum_{j=1}^V LCS_{\cup}(Q_i, A_j)}{n} \quad (8-3)$$

$$X_5 = \frac{(1 + \beta^2)R_{LCS}P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (9-3)$$

که در آن  $M$  تعداد پرسش‌ها در یک مکالمه و  $\beta = \frac{P_{LCS}}{R_{LCS}}$  خواهد بود. در این ویژگی  $Q_i$  شامل  $U$  جمله با  $P$  کلمه و مجموعه جواب‌ها شامل  $V$  جمله با  $n$  کلمه است.

### - ویژگی هشتم:

در این ویژگی، بزرگ‌ترین زیر رشته مشترک بین هر پرسش و مجموعه پاسخ‌ها را یافته و درون یک مجموعه قرار داده، سپس در بین همه اعضای این مجموعه بزرگ‌ترین زیر رشته حاصل شده را انتخاب می‌نماییم. روابط به صورت زیر تعریف گردید:

$$R_{LCS} = \frac{1}{N} \times \sum_{i=1}^N \max \left( \sum_{j=1}^P \frac{LCS(Q_i, A_j)}{L_{Q_i}} \right) \quad (10-3)$$

$$P_{LCS} = \frac{1}{N} \times \sum_{i=1}^N \max \left( \sum_{j=1}^P \frac{LCS(Q_i, A_j)}{L_{A_j}} \right) \quad (11-3)$$

$$X_6 = \frac{(1 + \beta^2)R_{LCS}P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (12-3)$$

که در این رابطه  $\beta = 1$ ،  $P$  تعداد جواب‌ها و  $N$  تعداد پرسش‌ها است.

### - ویژگی نهم:

برای محاسبه امتیاز هر مکالمه معادله ۳-۱۵ پیشنهاد گردید. در این ویژگی فرض گردید که دو مجموعه  $S_i$  و  $S_j$  داریم که  $S_i$  از  $N$  جمله با  $K$  کلمه و  $S_j$  با  $P$  جمله با  $T$  کلمه می‌باشند. بنابراین با به‌روزرسانی روابط قبلی، روابط ذیل حاصل گردید:

$$R_{LCS} = \frac{1}{N} \times \sum_{S_i \in S_1} \max_{S_j \in S_2} (LSC(S_i, S_j)) \quad (13-3)$$

$$P_{LCS} = \frac{1}{P} \times \sum_{S_i \in S_1} \max_{S_j \in S_2} (LSC(S_i, S_j)) \quad (14-3)$$

$$X_7 = \frac{(1 + \beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (15-3)$$

که در رابطه ۳-۱۵ مقدار  $\beta = 1$  در نظر گرفته شد. همچنین مجموعه‌های  $S_1$  و  $S_2$  را به صورت (  $Q_i, Q_{i+1}$ ),  $(Q_i, A_{i+1})$ ,  $(Q_{i+1}, A_i)$ ,  $(Q_i, A_i)$ ,  $(A_i, A_{i+1})$  در نظر گرفتیم. بنابراین به ازای هر مجموعه مقدار این ویژگی نیز محاسبه گردید.

### - ویژگی هشتم:

در این ویژگی پیشنهاد شده، N-gram‌های مشترک بین مجموعه پرسش‌ها و پاسخ‌ها را به دست آورده و بعد از نرمال‌سازی به عنوان امتیاز یک مکالمه گزارش گردید. در این ویژگی، با فرض در اختیار داشتن دو مجموعه از سؤالات و جواب‌ها و بر اساس روابط زیر، میزان امتیاز هر  $Q_i$  با مجموعه جواب‌ها را محاسبه و در نهایت توسط رابطه ۳-۱۸ امتیاز هر مکالمه محاسبه گردید.

$$R_{skip_n} = \frac{1}{t} \times \frac{1}{k} \times \sum_{i=1}^t \sum_{j=1}^k \frac{skip_n(Q_i, A_j)}{C(m, n)} \quad (16-3)$$

$$P_{skip_n} = \frac{1}{t} \times \frac{1}{k} \times \sum_{i=1}^t \sum_{j=1}^k \frac{skip_n(Q_i, A_j)}{C(L, n)} \quad (17-3)$$

$$X_8 = \frac{1 + \beta^2 \times R_{skip_n} \times P_{skip_n}}{R_{skip_n} + \beta^2 \times P_{skip_n}} \quad (18-3)$$

که در آن  $t$  تعداد سؤالات،  $k$  تعداد پاسخ‌ها،  $n$  اندازه N-gram (در اینجا مقدار ۲ و ۳ در نظر گرفته شد و skip همان N-gram در رابطه است)،  $m$  طول سؤال  $Q_i$ ،  $L$  طول پاسخ  $A_j$ ،  $C$  ترکیب و  $\beta = 1$  در نظر گرفته شد.

## - ویژگی نهم (امتیازدهی به جملات):

در این ویژگی، یک جفت پرسش و پاسخ را به صورت یک جمله در نظر گرفته و امتیاز آن را توسط رابطه ۳-۲۰ محاسبه میکنیم. سپس با توجه به امتیازات بدست آمده برای هر جمله، امتیاز نهایی برای یک مکالمه محاسبه گردید. نحوه امتیازدهی به کلمات و جملات بدین صورت است که ابتدا امتیاز مربوط به کلمات را محاسبه و سپس بر اساس امتیاز به دست آمده برای کلمات با استفاده از رابطه ۳-۲۰، امتیاز هر جمله محاسبه می‌شود.

$$Word\_score = K \times f_{word} \quad (19-3)$$

$$Sentence\_score = \sum Word\_score \quad (20-3)$$

که در آن  $K$  یک عدد ثابت و  $f_{word}$  تعداد تکرار کلمه در متن است. از طرفی با توجه به موقعیت مکانی هر جمله، امتیاز متفاوتی به آن تخصیص داده خواهد شد. در این امتیازدهی با توجه به اینکه معمولاً در هر مکالمه، جملات میانی دارای ارزش اطلاعاتی بالاتری نسبت به جملات ابتدایی و پایانی هر مکالمه هستند (بر اساس مجموعه داده تهیه شده این فرض صورت پذیرفت). بنابراین بر اساس موقعیت هر یک از جملات ارزش‌گذاری برای هر جمله به صورت زیر پیشنهاد گردید.

$$P_{score_i} = \begin{cases} 1 - \frac{n-i+1}{n} & i \leq \frac{n}{2} \\ 1 - \frac{i-3}{n} & \frac{n}{2} < i \leq n \end{cases} \quad (21-3)$$

که در آن  $i$  موقعیت هر جمله و  $n$  تعداد جملات هر مکالمه است. با توجه به ارزش هر جمله و امتیاز آن امتیاز نهایی یک مکالمه محاسبه می‌شود (رابطه ۳-۲۲).

$$X_9 = \frac{1}{N} \times \sum_{j=1}^N (Sentence\_score_j + P_{score_i}) \quad (22-3)$$

## - ویژگی دهم (فاصله همینگ):

یکی از ویژگی‌های مورد استفاده برای اندازه‌گیری تشابه بین جملات یک مکالمه فاصله همینگ می‌باشد. هر جمله در مکالمه، از تعدادی کلمه تشکیل شده است با به دست آوردن میزان شباهت بین دو کلمه، میزان شباهت بین جملات را می‌توان محاسبه نمود. فاصله همینگ دو کلمه برابر با تعداد حروف متناظر نامتشابه است. بنابراین این معیار میزان تفاوت بین دو کلمه را نشان می‌دهد. برای محاسبه شباهت عدد حاصل را بر طول کلمات تقسیم و از عدد یک کسر می‌نماییم. این کار را برای کل کلمات جمله انجام داده و در نهایت بر اساس اعداد به دست آمده میزان شباهت بین دو جمله محاسبه می‌شود. به‌طور مثال فرض کنید جمله  $Q_i$  شامل کلمات  $w_1 w_2 w_3 w_4 w_5$  با طول  $m$  و پاسخ  $A_i$  شامل  $w_1 w_2 w_6 w_7 w_8$  با طول  $n$  باشد. همچنین  $m$  و  $n$  تعداد حالت‌های ممکن ترکیب باشند. بنابراین معیار شباهت بین دو کلمه و یک جمله به صورت زیر محاسبه می‌شود.

$$Similarity_{words} = 1 - \frac{Hamming\_Distance(A,B)}{Max(|A|,|B|)} \quad (23-3)$$

$$X_{10} = \frac{1}{C(m,n)} \times \sum_{j=1}^{C(m,n)} Similarity\_words_j \quad (24-3)$$

## - ویژگی یازدهم (محاسبه امتیاز کلمات):

به دلیل اینکه معمولاً، یک مکالمه صورت گرفته در سیستم‌های پرسش و پاسخ تعاملی درباره موضوع خاصی بین کاربر و سیستم صورت می‌پذیرد، این باور وجود دارد که طرفین مکالمه از واژگان معینی برای ادامه بحث یا تشریح دقیق جنبه‌های مختلف موضوع استفاده و یا از تکرار آن‌ها استفاده می‌کنند. بنابراین در این ویژگی برای هر مکالمه گراف هم‌رخداد<sup>1</sup> کلمات ترسیم می‌شود. ویژگی این گراف در این است که کلمات هم‌رخداد در متن بدون استفاده از اندازه پنجره مشخص تعیین می‌شوند. با در نظر گرفتن تعداد رخداد هر واژه به عنوان عامل تعیین درجه اهمیت واژگان، در این گراف، تعداد تکرار

<sup>1</sup> Co-occurrence



هر کلمه در مکالمه و اینکه هر کلمه با چه کلمه دیگری آمده، نمایش داده می‌شود. با توجه به گراف حاصل، از روی آن، فرکانس کلمه، درجه کلمه در گراف و نسبت درجه به فرکانس کلمه محاسبه می‌گردد. نسبت درجه به فرکانس را به عنوان امتیاز نهایی هر کلمه مشخص کرده مجموع امتیازات هر کلمه به عنوان امتیاز هر مکالمه در نظر گرفته می‌شود.

$$Score_{w_i} = \frac{\deg(w)}{freq(w)} \quad (25-3)$$

$$X_{11} = \frac{1}{N} \times \sum_{j=1}^N Score_{w_j} \quad (26-3)$$

### - ویژگی دوازدهم (Tf-Idf):

پارامتر TF-Idf یکی از ویژگی‌های آماری است که بر اساس آن می‌توان در سیستم‌های بازیابی اطلاعات، میزان شباهت میان کلمات منتخب با یک سند را محاسبه کرد. در این سیستم‌ها، از این ویژگی برای محاسبه میزان شباهت هر سند با پرسش مطرح شده استفاده می‌شود. در روش پیشنهادی، برای این ویژگی میزان تکرار یک کلمه در مکالمه، در مقابل تعداد تکرار آن در کل مکالمات، محاسبه می‌شود. سپس مجموع این امتیازات برای کلمات یک مکالمه محاسبه و پس از نرمال‌سازی به عنوان امتیاز یک مکالمه گزارش می‌شود. رابطه زیر نحوه محاسبه وزن هر کلمه حاصل را نشان می‌دهد.

$$X_{12} = F_w \times \log \frac{N_{sentence}}{NC_{sentence}} \quad (27-3)$$

که در آن  $F_w$  تعداد تکرار هر کلمه در جملات یک مکالمه،  $N_{sentence}$  تعداد جملات یک مکالمه و  $NC_{sentence}$  تعداد جملاتی است که شامل کلمه نام است.

### - ویژگی سیزدهم تا شانزدهم:

یکی دیگر از ویژگی‌های پیشنهادی استفاده از معیار شباهت بین سؤال و جواب هر مکالمه است که با استفاده از روابط ۲۸-۳ تا ۳۱-۳ بدست می‌آید. در این ویژگی، دو مجموعه  $S_1$  و  $S_2$  به ترتیب

مجموعه N-gram های مربوط به سؤال و جواب می‌باشند. بنابراین مقدار شباهت بین این دو مجموعه با استفاده از رابطه ۳-۲۸ به صورت جداگانه برای هر جفت پرسش و پاسخ محاسبه گردید و میانگین امتیاز به دست آمده برای هر مکالمه محاسبه و به عنوان یک ویژگی در ماتریس ویژگی‌ها ذخیره گردید.

$$X_{13} = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \quad (28-3)$$

نحوه محاسبه ویژگی ۱۵، ۱۴، ۱۶ همانند ویژگی سیزدهم است. اما معیار در نظر گرفته شده برای هر یک به ترتیب از روابط زیر استفاده گردید.

$$X_{14} = \frac{|S_1 \cap S_2|}{\sqrt{|S_1| \times |S_2|}} \quad (29-3)$$

$$X_{15} = \frac{|S_1 \cap S_2|}{|S_2|} \quad (30-3)$$

$$X_{16} = \frac{|S_2|}{|S_1|} \quad (31-3)$$

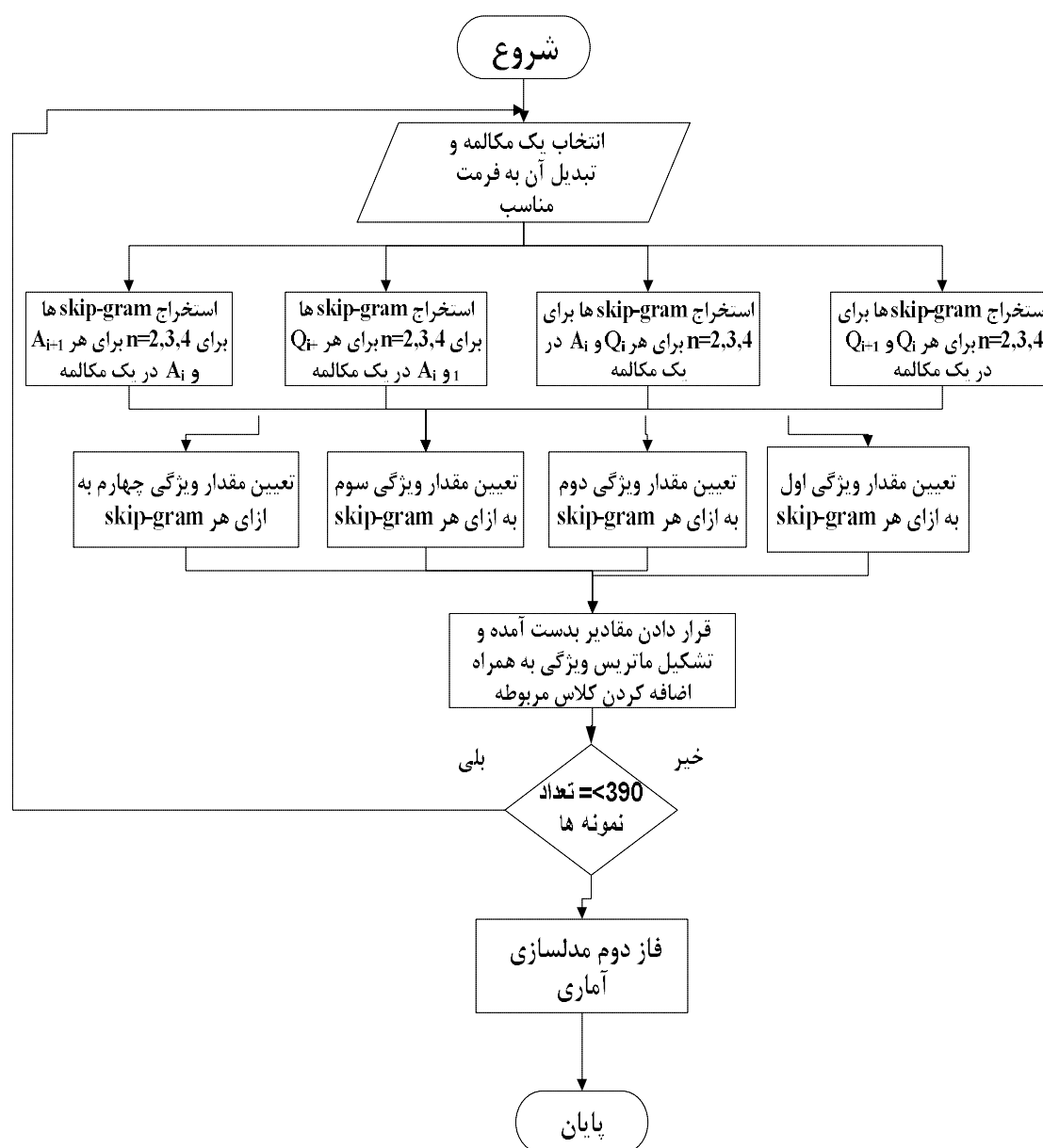
مجموعه  $S_1$  و  $S_2$  مجموعه N-gram های مربوط به سؤال یا جواب بسته به حالت در نظر گرفته شده (منظور مجموعه‌های در نظر گرفته است مثلاً زوج پرسش و پاسخ یا مجموعه پرسش‌ها و یا مجموعه پاسخ‌ها) می‌باشند. مقدار این ویژگی برای N-gram، ۲ تا ۳ محاسبه شد. بنابراین با استفاده از روابط ۳-۲۸ تا ۳-۳۱، در مجموع هشت ویژگی بدست آمد. در راستای رسیدن به مدل مناسب، چندین آزمایش مختلف با استفاده از روش‌های مرسوم در تحلیل داده مانند رگرسیون صورت پذیرفت. همچنین با توجه به عدم حدس آگاهانه نسبت به معادله خط منطبق بر ویژگی‌ها بدلیل تعدد ویژگی‌های تعریف شده، مجبور به استفاده از چندین رگرسیون مختلف در این راه شدیم که برای حل این معضل از روش برنامه‌نویسی بیان ژن بهره گرفته شد. نتایج حاصل هر یک از این روش‌ها در فصل آتی گزارش شده است. همچنین یکی از ویژگی‌هایی که در ارزیابی مورد توجه بسیاری از ارزیاب‌ها

قرار می‌گرفت، انسجام بین پرسش و پاسخ‌های رد و بدل شده بین کاربران و سیستم می‌باشد. بنابراین چهار ویژگی که در تعیین انسجام بکار گرفته شدند، توسط یک آزمون آماری به اثبات عملکرد آنها در تعیین انسجام پرداختیم.

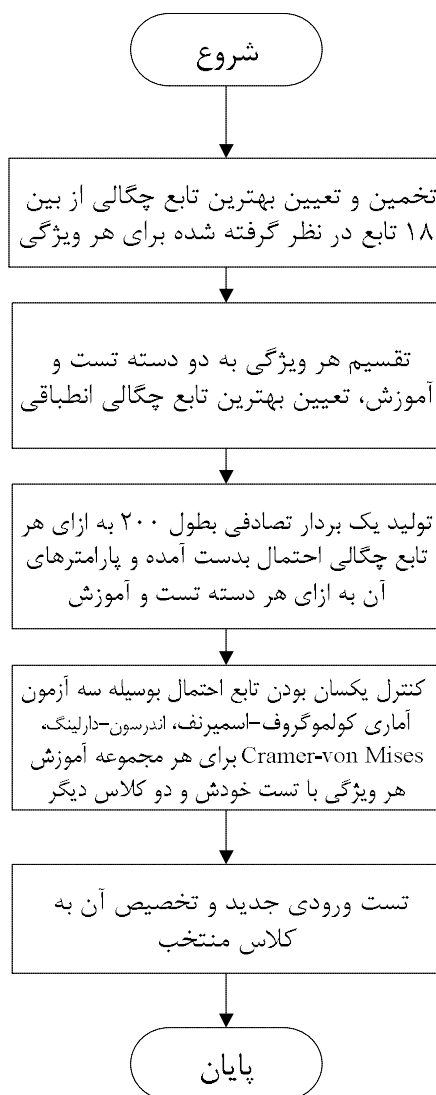
### ۳-۴- تعیین انسجام متن در سیستم‌های IQA با ارائه یک مدل احتمالاتی

با توجه به اینکه در N-gramها، برای مقادیر بیشتر از ۲ برای n انسجام محلی وجود دارد [۵۳]. ما از این نکته بهره گرفته و از آن برای مدل‌سازی تعیین انسجام در متن خروجی یک سیستم IQA بهره گرفتیم. برای اندازه‌گیری انسجام محلی و رسیدن به انسجام کلی ما از چهار رابطه ۳-۲۸ تا ۳-۳۱ استفاده کردیم و از این چهار رابطه توانستیم ۴۸ ویژگی ایجاد نمودیم. در این روابط، دو مجموعه  $S_1$  و  $S_2$  تعریف شد. به عنوان مثال، برای یک جفت پرسش و پاسخ،  $S_1$  مجموعه N-gramهای مربوط به پرسش و  $S_2$  مجموعه N-gramهای مربوط به پاسخ در نظر گرفته شد. همانطور که می‌دانیم، یک مکالمه از تعدادی پرسش و پاسخ تشکیل شده است. بنابراین مقدار هر یک از این روابط به صورت جداگانه برای هر جفت پرسش و پاسخ محاسبه و در نهایت میانگین مقادیر بدست آمده برای هر مکالمه محاسبه و در ماتریس ویژگی مربوط به آن مکالمه قرار گرفت. برای تشکیل ماتریس ویژگی با استفاده از این روابط، یکبار برای جفت  $(A_i - Q_i)$  سپس برای جفت‌های  $(Q_{i+1} - Q_i)$ ،  $(Q_{i+1} - A_i)$  و  $(A_{i+1} - Q_i)$  در محاسبات استفاده شد. که در آن  $A_i$ ها جواب‌های یک مکالمه و  $Q_i$ ها سؤالات یک مکالمه است. مقدار  $N$  ۲ تا ۴ فرض گردید. بنابراین با توجه به چهار گروه در نظر گرفته شده، مقدار  $N$  و چهار رابطه تعریف شده، ۴۸ ویژگی حاصل شد. با توجه به تعداد مکالمات، اندازه ماتریس ویژگی  $390 \times 49$  حاصل گردید که ۴۸ ستون آن مربوط به ویژگی‌ها و ستون آخر مربوط به کلاس مکالمه است. مکالمه‌های موجود در پایگاه داده به صورت دستی به سه کلاس تقسیم گردید. در کلاس اول مکالمات منسجم، در کلاس دوم مکالمات نیمه منسجم و در کلاس سوم مکالمات غیرمنسجم قرار گرفته‌اند. دسته‌بندی هر مکالمه بر اساس امتیازات داده شده کاربران به انسجام مکالمه و با نظارت فرد

خبره انسانی صورت پذیرفت. تعداد مکالمه‌های هر کلاس ۱۳۰ نمونه بود که با توجه به سه دسته موجود، در مجموع ۳۹۰ نمونه در نظر گرفته شد. یک الگوریتم مناسب برای محاسبه انسجام باید قادر باشد مجموعه تفاوت‌های هر مکالمه را با مکالمه‌های دیگر به صورت کاملاً پرننگ دیده و در همان حال، مجموعه تفاوت‌های هر مکالمه با خودش را در حد امکان از بین ببرد. شکل ۲-۳ ساختار روش در نظر گرفته شده برای اثبات کارایی ویژگی‌های تعریف شده در تعیین انسجام یک مکالمه حاصل از تعامل یک کاربر با یک سیستم IQA را در دو فاز نمایش می‌دهد.



الف) مرحله اول



ب) مرحله دوم

شکل ۳-۲ تشریح مراحل روش پیشنهادی برای تعیین انسجام متن در سیستم‌های IQA

نتایج حاصل از اثبات کارایی این ویژگی‌ها برای تعیین انسجام یک مکالمه در سیستم‌های IQA در فصل آتی گزارش شده است.

### ۳-۵ مدل‌سازی داده‌ها با استفاده از آنالیز رگرسیونی

مدل‌سازی داده‌های غیرخطی یک کار معمول در علم اطلاعات و حوزه تحلیل داده‌ها است. به ندرت فرآیند طبیعی را می‌توان یافت که نتیجه آن به طور خطی با متغیرهای مستقل مربوط به یک

مجموعه از داده‌ها تغییر نماید. بنابراین، به یک روش ساده و قوی نیاز هست تا به سرعت داده‌های اندازه‌گیری شده را در برابر مجموعه‌ای از متغیرها (که داده‌های اندازه‌گیری شده می‌توانند یک عملکرد غیرخطی پیچیده داشته باشند) مدل سازی نماید. بنابراین باید یک ابزار نسبتاً رایج و قابل دسترس برای مهندسی یادگیری ماشین در نظر گرفت که قابلیت تعمیم به موضوعات دیگر را داشته باشد. رگرسیون اولین ابزار جستجو برای یافتن مدل می‌باشد. رگرسیون در لغت به معنی بازگشت و در علم آمار، به عنوان یک نوع رابطه یا تابع ریاضی بین متغیر وابسته از یک طرف و متغیرهای مستقل از سوی دیگر تعریف می‌شود. در مدل‌های آماری، تحلیل رگرسیون یک فرآیند آماری برای تخمین روابط بین متغیرها می‌باشد. این روش شامل تکنیک‌های زیادی برای مدل‌سازی و تحلیل متغیرهای خاص و منحصر بفرد، با تمرکز بر رابطه بین متغیر وابسته و یک یا چند متغیر مستقل، می‌باشد. در استفاده از تحلیل رگرسیونی چند فایده وجود دارد که از جمله می‌توان به موارد ذیل اشاره نمود:

الف- امکان نشان دادن رابطه معنی‌دار بین متغیر مستقل و وابسته وجود دارد. برای این رابطه معنی‌دار، گاهی از همبستگی استفاده می‌شود بنابراین باید به این نکته توجه داشت که رگرسیون دارای ماهیت پیش‌بینی میزان متغیر وابسته از روی مدل (مدلی که متغیر مستقل یکی از مؤلفه‌های آن می‌باشد) است درحالی‌که همبستگی صرفاً نشان‌دهنده شدت رابطه متغیر (متغیرهای) مستقل و وابسته است.

ب- رگرسیون نشان دهنده میزان تاثیر متغیرهای مستقل بر وابسته می‌باشد به عبارتی دیگر، چگونه مقدار متغیر وابسته با تغییر هرکدام از متغیرهای مستقل و با ثابت بودن دیگر متغیرهای مستقل، تغییر می‌کند.

ج- امکان مقایسه اثرات متغیرهای اندازه‌گیری شده در مقیاس‌های مختلف را فراهم می‌آورد. و موارد دیگری که این امکان را فراهم می‌سازند تا بهترین مجموعه متغیرهایی که برای ساختن مدل‌های پیش‌بینی استفاده می‌شود را انتخاب، حذف و ارزیابی نمایند. در تحلیل رگرسیون، تعیین پراکندگی متغیر وابسته اطراف تابع رگرسیون مورد توجه است که می‌تواند توسط یک توزیع احتمال

توضیح داده شود. تحلیل رگرسیون به صورت گسترده برای پیش بینی استفاده شده است. در شرایط خاصی این تحلیل، برای استنتاج روابط عالی بین متغیرهای مستقل و وابسته می‌تواند استفاده شود. هر چند این می‌تواند موجب روابط اشتباه شود. بنابراین در استفاده برای این موارد باید دقت کافی داشت. انواع مختلفی از تکنیک‌های رگرسیونی برای پیش‌بینی مطرح شده‌اند. این تکنیک‌ها، در بیشتر موارد توسط سه معیار تعداد متغیرهای مستقل، نوع متغیرهای وابسته و شکل خط رگرسیون انتخاب می‌شوند. همچنین این امکان وجود دارد که هر فرد خلاق بتواند یک رگرسیون جدید پیشنهاد نماید در صورتیکه احساس نماید در مسئله خود نیاز به استفاده از ترکیبی از سه معیار اشاره شده دارد که افراد قبلاً از آن استفاده نکرده‌اند. رگرسیون خطی و لجستیک معمولاً اولین الگوریتم‌هایی هستند که در پیش‌بینی مدل از طرف کاربران بیشتر مورد استفاده قرار می‌گیرد در حالی که واقعیت این است که تعداد بیشماری روش رگرسیونی برای پیش‌بینی وجود دارد که به تنهایی هر یک از آن‌ها دارای اهمیت بوده و متناسب با شرایط مسئله به کار گرفته می‌شوند که از جمله این روش‌ها می‌توان به رگرسیون خطی<sup>۱</sup>، چندجمله‌ای<sup>۲</sup>، ریج (مضرس)<sup>۳</sup>، لجستیک<sup>۴</sup>، چندمرحله‌ای<sup>۵</sup>، لسو<sup>۶</sup>، Elasticnet و دیگر موارد اشاره نمود (برای آشنایی بیشتر پیوست الف مطالعه شود).

با توجه به تعدد توابع رگرسیونی، در انتخاب مدل مناسب با چالش‌ها و معضلات زیادی مواجه هستیم. برای درک بهتر موضوع، به شکل ۳-۳ توجه نمایید. همانطور که در قسمت الف و ب نشان داده شده، زمانی که ما تنها یک ویژگی و یا دو ویژگی داریم به راحتی می‌توانیم یک حدس آگاهانه نسبت به مدل منطبق بر داده‌ها داشته باشیم. اما این سوال پیش می‌آید که آیا برای داده‌ها با ابعاد بالاتر این امکان نیز وجود دارد (قسمت ج شکل)؟. در پاسخ به این سوال می‌توان گفت، در صورتی که ابعاد داده‌ها

<sup>1</sup> Linear regression

<sup>2</sup> Polynomial regression

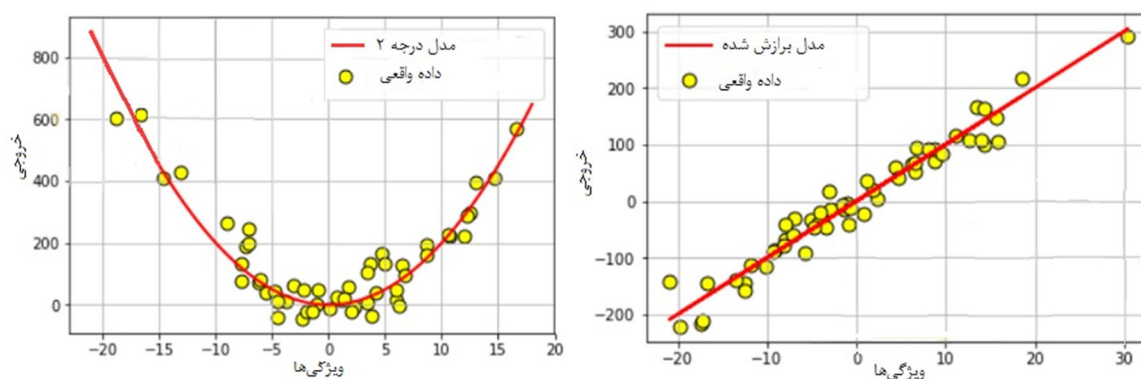
<sup>3</sup> Ridge regression

<sup>4</sup> Logistic regression

<sup>5</sup> Stepwise regression

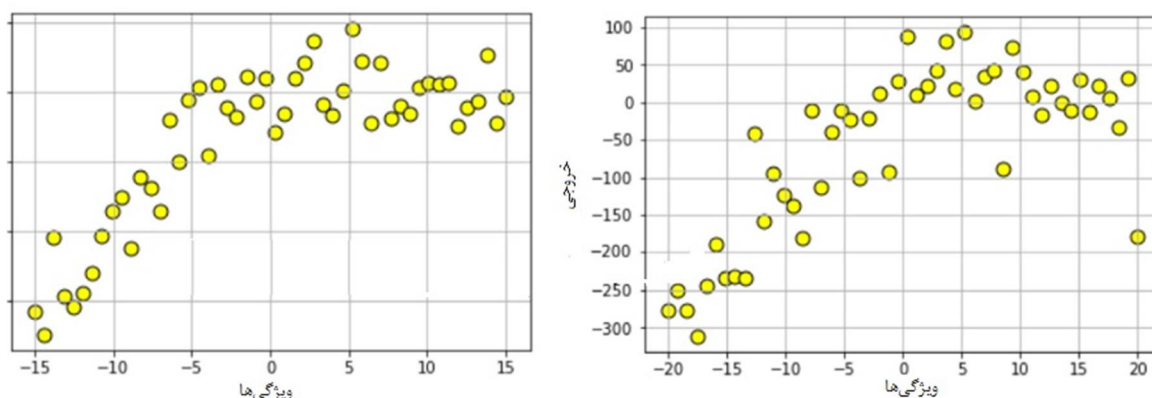
<sup>6</sup> Lasso regression

بیشتر از سه شود، دیگر امکان حدس زدن تابع به راحتی وجود نداشته و در این حالت هر یک از ویژگی‌ها بر روی ویژگی‌های دیگر تاثیر خواهند گذاشت.



(ب) مدل درجه دو برازش شده

(الف) مدل خطی برازش شده



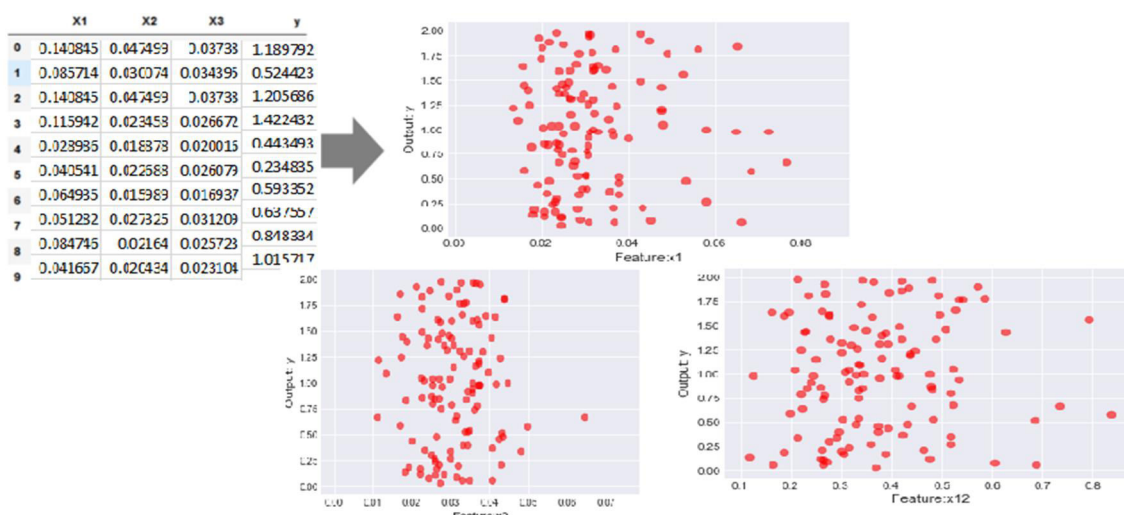
(ج) مقایسه همزمان چند ویژگی جهت انتخاب مدل

شکل ۳-۳ بررسی روند انتخاب نوع مدل رگرسیونی با توجه به متغیرهای ورودی

بنابراین نیاز به یک ابزار مهندسی برای رسیدن به معادله خط مناسب ضروری به نظر می‌رسد. رگرسیون خطی اولین ابزار جستجو برای یافتن مدل می‌باشد. توجه داشته باشید که "LINEAR" در مدل رگرسیون خطی به ضرایب اشاره دارد و نه به درجه ویژگی‌ها، بنابراین ویژگی‌ها (یا متغیرهای مستقل) می‌توانند از هر درجه یا حتی توابع مختلف مانند لگاریتمی، سینوسی و غیره باشند و با استفاده از این توابع و مدل خطی، یک مدل از نمونه‌های تقریبی مدل سازی شود. بنابراین در روش پیشنهادی، برای یافتن بهترین مدل، انواع مختلف رگرسیون توسط نرم افزارهای مختلف مانند متلب،



پایتون و<sup>۱</sup> SPSS مورد آزمایش و تحلیل قرار گرفت تا بهترین مدل انتخاب گردد. با توجه به اینکه تعداد ویژگی‌های استخراج شده زیاد است و امکان حدس آگاهانه مدل به راحتی امکان‌پذیر نیست (شکل ۳-۴). بنابراین نیاز به تست مدل‌های مختلف برای رسیدن به مدل مناسب، براساس معیارهای تعیین شده می‌باشیم. مقادیر تعدادی از ویژگی‌ها در شکل ۳-۴ نشان داده شده است اما، همانطور که انتظار می‌رفت، با توجه به تعداد زیاد ویژگی‌ها و حدس در مورد تاثیر هر ویژگی بر یکدیگر و نیز خروجی به راحتی امکان‌پذیر نمی‌باشد. لذا مشاهده آن‌ها کمک زیادی در پیش‌بینی مدل نمی‌کند. بنابراین، ما تصمیم گرفتیم که مدل‌های متفاوت اعم از خطی و غیرخطی با درجه‌های مختلف چند درجه‌ای را بر روی ماتریس ویژگی تست نموده تا بهترین مدل حاصل گردد. اما چگونه می‌توان یک مدل قوی ایجاد نمود که همه این شرایط را در کنار هم داشته باشد. یک پروژه یادگیری ماشینی یک کار مدل‌سازی واحد نیست. در رایج‌ترین شکل خود، شامل تولید داده، تمیزکردن داده و تبدیل آن، مدل مناسب، اعتبارسنجی مدل، تست صحت مدل و کاربرد نهایی است. بنابراین باید به همه این عوامل در زمان شبیه‌سازی دقت نمود. همچنین یکی از نکات مهمی که در شرایط شبیه‌سازی باید به آن دقت نموده و از رخ دادن آن باید جلوگیری نمود، بیش‌برازش<sup>۲</sup> می‌باشد.



شکل ۳-۴ نمایی قسمتی از ماتریس ویژگی به همراه نمودار پراکندگی برخی از ویژگی‌ها جهت نمایش عدم امکان حدس تابع

<sup>1</sup> Statistical Package for the Social Science

<sup>2</sup> Over fitting

وقتی در یک مدل، داده‌های آموزشی، شامل ویژگی‌هایی برخواسته از نویز یا داده‌هایی با واریانس بالا ایجاد گردد، باعث به وجود آمدن پدیده بیش‌برازش می‌شود که این امر سبب شده تا مدل بیش از حد به داده‌های آموزشی وابسته گردد و نتواند توزیعی را که داده‌های واقعی از آن تولید شده بودند را مدل کند. همچنین این موضوع سبب می‌گردد تا صحت مدل در پیش‌بینی داده‌هایی که در نمونه‌گیری وجود ندارند به شدت کاهش یابد. در واقعیت، بین اندازه فضای، مدل‌هایی که یک یادگیرنده می‌تواند تولید نماید و همچنین ریسک ایجاد بیش‌برازش، یک رابطه مستقیم وجود دارد. یعنی هر چقدر فضای بین مدل‌هایی که یک یادگیرنده می‌تواند انتخاب کند افزایش یابد، ریسک به وجود آمدن بیش‌برازش نیز بیشتر می‌شود. به عبارت دیگر، مدل بدست آمده، بالقوه می‌تواند بر توزیع داده‌ها منطبق گردد. این خاصیت را می‌توان به صورت حالتی خاص، از توازن انحراف و واریانس نیز بیان نمود. از دلایل بروز بیش‌برازش می‌توان به وجود نویز زیاد در داده‌های آموزشی و تعداد کم نمونه‌ها نیز اشاره کرد. همچنین برای پرهیز و حل آن دو روش و دید کلی وجود دارد. راه‌حل اول، می‌توان به جلوگیری از رشد درخت قبل از رسیدن به مرحله‌ای که بطور کامل داده‌های آموزشی را دسته‌بندی نماید و دیگری رشد کامل درخت و سپس هرس کردن شاخه‌هایی که مفید نیستند، اشاره نمود. بنابراین باید در شرایط شبیه‌سازی برای رسیدن به مدل مناسب به این پدیده نیز توجه داشت. با توجه به اینکه تعداد نمونه‌های ذخیره شده برای مکالمات برابر با ۵۴۰ نمونه می‌باشد. از تصدیق تقاطعی<sup>۱</sup> برای ایجاد مجموعه آزمون و آموزش به نسبت ۷۰ درصد برای آموزش و ۳۰ درصد برای تست بهره گرفتیم. با توجه به اینکه هدف ارائه یک مدل آماری جهت ارزیابی سیستم‌های IQA است و با توجه به نظرات موجود کاربران در تعامل با این سیستم‌ها و استخراج ویژگی‌ها از روی متن‌های تولید شده، به دنبال استفاده از رگرسیون برای پیش‌بینی این نظرات هستیم. در مرحله اول آزمایش ویژگی‌های استخراج شده به عنوان متغیرهای مستقل و نظرات انسانی به عنوان متغیر وابسته در نظر گرفته شد. سپس جهت انتخاب نوع درست از انواع رگرسیون، مدل‌های مختلف مورد بررسی قرار گرفت تا بهترین مدل

---

<sup>۱</sup> Cross validation

انتخاب گردد. با توجه به عدم اطلاع در مورد بهترین مدل و تعداد بالای ویژگی‌های استخراج شده، از رگرسیون‌های مختلف مانند رگرسیون خطی چندگانه، ریج، لسو، توانی، نمایی، لگاریتمی و دیگر موارد رگرسیون استفاده گردید تا بهترین مدل انتخاب گردد. در فصل آتی نحوه انجام این کار گزارش شده است.

### ۳-۶ برنامه‌نویسی بیان ژن جهت تعیین مدل مناسب

یکی از مشکلات موجود برای به دست آوردن بهترین مدل رگرسیونی، داشتن یک حدس آگاهانه نسبت به مدل است. با توجه به مقادیر متفاوت ویژگی‌ها و تعداد بالای آن‌ها، امکان یافتن بهترین مدل با توجه به معیار ارزیابی معرفی شده و تست نمودن معادلات مختلف کاری طاقت فرسا است. لذا با توجه به قابلیت‌های برنامه‌نویسی بیان ژن که تکامل یافته برنامه‌نویسی ژنتیک است برای استخراج بهترین مدل استفاده نمودیم. برنامه نویسی ژنتیک<sup>۱</sup> نمونه‌ای از الگوریتم تکاملی و متعلق به حوزه هوش محاسباتی است. این نوع از برنامه‌نویسی در جهت خودکارسازی فرآیندها توسعه داده شده است [۵۵]. در GP اندازه کروموزوم‌ها به وسیله عملگرهای ژنتیک اصلاح شده، تغییر می‌کند. فضای جستجو در برنامه نویسی ژنتیک نه تنها شامل فضای مسئله است بلکه فضای نمایش را نیز در بر می‌گیرد و تفاوتی بین فضای جستجو و فضای نمایش قائل نمی‌شود. فضای جستجوی برنامه‌نویسی ژنتیک، عبارت است از فضای جستجوی همه ترکیب‌های ممکن از سمبل‌های اولیه‌ای که در مسئله از آن‌ها ساخته می‌شود. برنامه‌ها در قالب یک درخت یا فرم خطی ارائه شده و از تلفیق به عنوان عملگر اصلی برنامه‌نویسی ژنتیک استفاده می‌شود. در واقع زیر درخت‌های مربوطه به والدین را بدون اینکه خدش‌ای در اجرای برنامه ایجاد شود، با یکدیگر مبادله می‌کنند. همچنین عملیات جهش، یک زیر درخت تصادفی ایجاد کرده و آن را جایگزین می‌کند. برنامه تا جایی افزایش می‌یابد تا درخت به ماکزیمم عمق خود برسد. برنامه نویسی بیان ژن<sup>۲</sup> تعمیم یافته الگوریتم ژنتیک می‌باشد. به این ترتیب

<sup>۱</sup> Genetic Programming (GP)

<sup>۲</sup> Gene Expression Programming (GEP)

که جمعیتی در جهت تکامل به صورت انتخابی شکل می‌دهد. سپس، جمعیت نامناسب را رها کرده و فرزندان اصلاح شده ایجاد می‌کند. در این روش، در ابتدای فرآیند، هیچگونه تابعی در نظر گرفته نشده و برنامه یک جمعیتی از افراد را مورد استفاده قرار داده و آن‌ها را مطابق برازندگی انتخاب می‌کند و تغییرات ژنتیکی را با استفاده از یک یا چند عملگر ژنتیکی اعمال می‌نماید. این روش قادر به بهینه سازی ساختار مدل و مؤلفه‌های آن می‌باشد. برنامه‌ریزی بیان ژن بر خلاف الگوریتم ژنتیک روی ساختار درختی فرمول‌ها به جای سلسله ارقام دودویی عمل می‌کند. ساختارهای درختی از مجموعه توابع (عملگرهای ریاضی مورد استفاده در فرمول‌ها) و ترمینال‌ها (متغیرهای مسئله و اعداد ثابت) ایجاد می‌شوند. قبل از مراحل اجرایی برنامه‌ریزی بیان ژن گام‌های مقدماتی زیر باید توسط کاربر تعیین شوند:

الف- مجموعه ترمینال‌ها (متغیرهای مسئله، اعداد ثابت تصادفی)

ب- مجموعه عملگرهای ریاضی مورد استفاده در فرمول‌ها

ج- انتخاب تابع برازش<sup>۱</sup> مناسب برای سنجش برازش فرمول‌ها

د- تعیین پارامترهای کنترل کننده اجرای برنامه (اندازه جمعیت، احتمال مربوط به کارگیری

عمل‌های ژنتیکی و جزئیات دیگر مربوط به اجرای برنامه)

ه- معیار پایان و ارائه نتایج اجرای برنامه (مثل تعداد تولید جمعیت جدید، تعیین یک مقدار

مشخص برای برازش فرمول‌ها که اگر میزان برازش برابر یا بیشتر از آن مقدار شد، اجرا متوقف شود) بنابراین می‌توان به این صورت بیان کرد که اولین مرحله در الگوریتم GEP تولید جمعیت اولیه از راه حل‌ها است. این موضوع می‌تواند به وسیله فرآیند تصادفی و یا از طریق اطلاعات موجود درباره مسئله انجام شود. سپس کروموزوم‌ها به صورت بیان درختی نشان داده شده، که این هم مطابق با یک تابع برازش ارزیابی می‌گردد. معمولاً تابع برازش به وسیله پردازش تعدادی تابع هدف<sup>۲</sup> (که مورد برازش نیز نامیده می‌شوند) ارزیابی می‌گردد. اگر متناسب با معیار ارزیابی در نظر گرفته شده، مقدار

<sup>۱</sup> Fitness Function

<sup>۲</sup> Cost Function

رضایت بخشی حاصل گردد و یا نسل‌ها به تعداد معینی برسند، تکامل متوقف می‌شود و بهترین راه حل یافت شده، گزارش داده می‌شود. از طرف دیگر اگر شرایط توقف یافت نشود، بهترین راه حل از نسل حاضر نگه داشته می‌شود (نخبه‌گزینی) و بقیه راه‌حل‌ها به فرآیند گزینشی<sup>۱</sup> واگذار می‌شوند. بر اساس گزینش، بهترین افراد، شانس بیشتری برای تولید فرزندان دارند. کل این روند برای چندین نسل تکرار می‌شود و با پیش رفتن نسل به جلو، انتظار می‌رود که کیفیت جمعیت نیز به‌طور متوسط بهبود یابد. (توضیحات تکمیلی در مورد برنامه‌نویسی بیان ژن در پیوست ب آورده شده است)

### ۳-۷ یادگیری عمیق

یکی از تکنیک‌هایی که امروزه در بحث مدل‌سازی مورد توجه محققین قرار گرفته است، استفاده از یادگیری عمیق می‌باشد. یادگیری عمیق زیر شاخه‌ای از یادگیری ماشین و بر مبنای مجموعه‌ای از الگوریتم‌ها است که مفاهیم انتزاعی سطح بالا در دادگان را مدل و این فرآیند را با استفاده از یک گراف عمیق مدل می‌کند. یادگیری عمیق دارای چندین لایه پردازشی، متشکل از چندین لایه تبدیلات خطی و غیرخطی است. این روش، در واقع از نوع عملکرد مغز انسان الهام گرفته و به ابزارهای پیشرفته مانند کارت‌های گرافیک قدرتمند برای محاسبات پیچیده و حجم زیادی از کلان داده نیاز دارد. حجم کم داده‌ها در این الگوریتم به نتایج و عملکردی ضعیف‌تر ختم می‌شود. برخلاف الگوریتم‌های یادگیری ماشین استاندارد که مسائل را به بخش‌های کوچک‌تر تقسیم و سپس آن‌ها را حل می‌کنند، یادگیری عمیق، مسائل را به شکل کامل حل می‌کند. هر چه حجم داده بیشتر و زمان بیشتری در اختیار الگوریتم‌های یادگیری عمیق قرار گیرد، نتیجه نهایی بهتر خواهد بود. یادگیری عمیق در واقع روشی نوین برای یادگیری همان شبکه‌های عصبی استاندارد چند لایه است. روش‌های مختلفی در یادگیری عمیق مانند شبکه‌های عصبی کانولوشن<sup>۲</sup>، شبکه باور عمیق<sup>۳</sup>، حافظه طولانی

<sup>1</sup> Selection

<sup>2</sup> Convolutional Neural Network (CNN)

<sup>3</sup> Deep Belief Network (DBN)

کوتاه مدت<sup>۱</sup> و دیگر موارد وجود دارد که هر کدام در حوزه‌های مختلف هوش مصنوعی کاربرد دارند. بزرگ‌ترین ویژگی LSTM امکان یادگیری وابستگی بلند مدت است که توسط شبکه‌های عصبی بازگشتی امکان‌پذیر نبود. برای پیش‌بینی گام زمانی بعدی نیاز است که مقادیر وزن‌ها در شبکه بروز رسانی شوند که این کار مستلزم حفظ اطلاعات گام‌های زمانی ابتدایی است. یک شبکه عصبی بازگشتی فقط می‌تواند تعداد محدودی از وابستگی‌های کوتاه مدت را یاد بگیرد، اما سری‌های زمانی بلند مدت مثل ۱۰۰۰ گام زمانی قابل یادگیری توسط RNNها نیستند. LSTMها می‌توانند این وابستگی‌های بلند مدت را به درستی یاد بگیرند (توضیحات بیشتر در مورد این شبکه در پیوست ج آورده شده است). در این تحقیق از شبکه LSTM برای مدل‌سازی مسئله پیش‌بینی امتیاز کاربر بر اساس چندین ورودی (مجموعه ویژگی‌ها) مورد استفاده قرار گرفت. نحوه چگونگی استفاده از این شبکه و نتایج حاصل از مدل‌سازی مسئله پیش‌بینی امتیاز داده شده توسط کاربر در فصل بعدی توضیح داده شده است.

### ۳-۸ معیار ارزیابی

برای ارزیابی نتایج حاصل از مدل به دست آمده با داده‌های واقعی، سه سنج آماری ضریب تعیین  $R^2$ ، جذر میانگین مربعات خطا (RMSE) و درصد میانگین مطلق خطا (MAPE) مورد استفاده قرار گرفت. روابط مربوط به این معیارها در زیر نشان داده شده است.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n}} \quad (3-32)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{e_i}{Y_i} \right| \times 100 \quad (3-33)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - y)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (3-34)$$

<sup>۱</sup> Long Short Term Memory (LSTM)

که در این روابط  $n$  تعداد پیش‌بینی‌ها و  $e_i$  خطای پیش‌بینی است که از تفاوت مقادیر پیش‌بینی شده و مقادیر واقعی به دست می‌آید و  $Y_i$  مقادیر واقعی،  $y$  مقدار پیش‌بینی شده و  $\bar{Y}$  میانگین مقادیر واقعی می‌باشد. مقدار  $R^2$  در واقع نشانگر میزان تطابق مقادیرهای پیش‌بینی شده با نتایج واقعی است. به عبارت دیگر  $R^2$  معیاری از قدرت پیش‌بینی مدل رگرسیون است. این معیار ضریب تعیین نیز نامیده می‌شود. محدوده تغییرات این کمیت بین صفر و یک است. به‌طور مثال اگر  $R^2 = 0.85$  باشد به این معنی است که ۸۵٪ کل تغییرات  $Y$  می‌تواند از طریق ارتباط خطی یا غیرخطی بین  $X$  و  $Y$  (از طریق معادله رگرسیون) تعریف شود و ۱۵٪ از متغیرهای  $Y$ ، تعریف نشده باقی می‌مانند. معیار RMSE مقدار پراکندگی مقادیرهای تجربی حول خط رگرسیون را نشان می‌دهد یا به عبارت دیگر میزان خطا را مشخص می‌نماید. پایین بودن مقدار RMSE نشان دهنده قدرت پیش‌بینی و اعتبار مدل است.





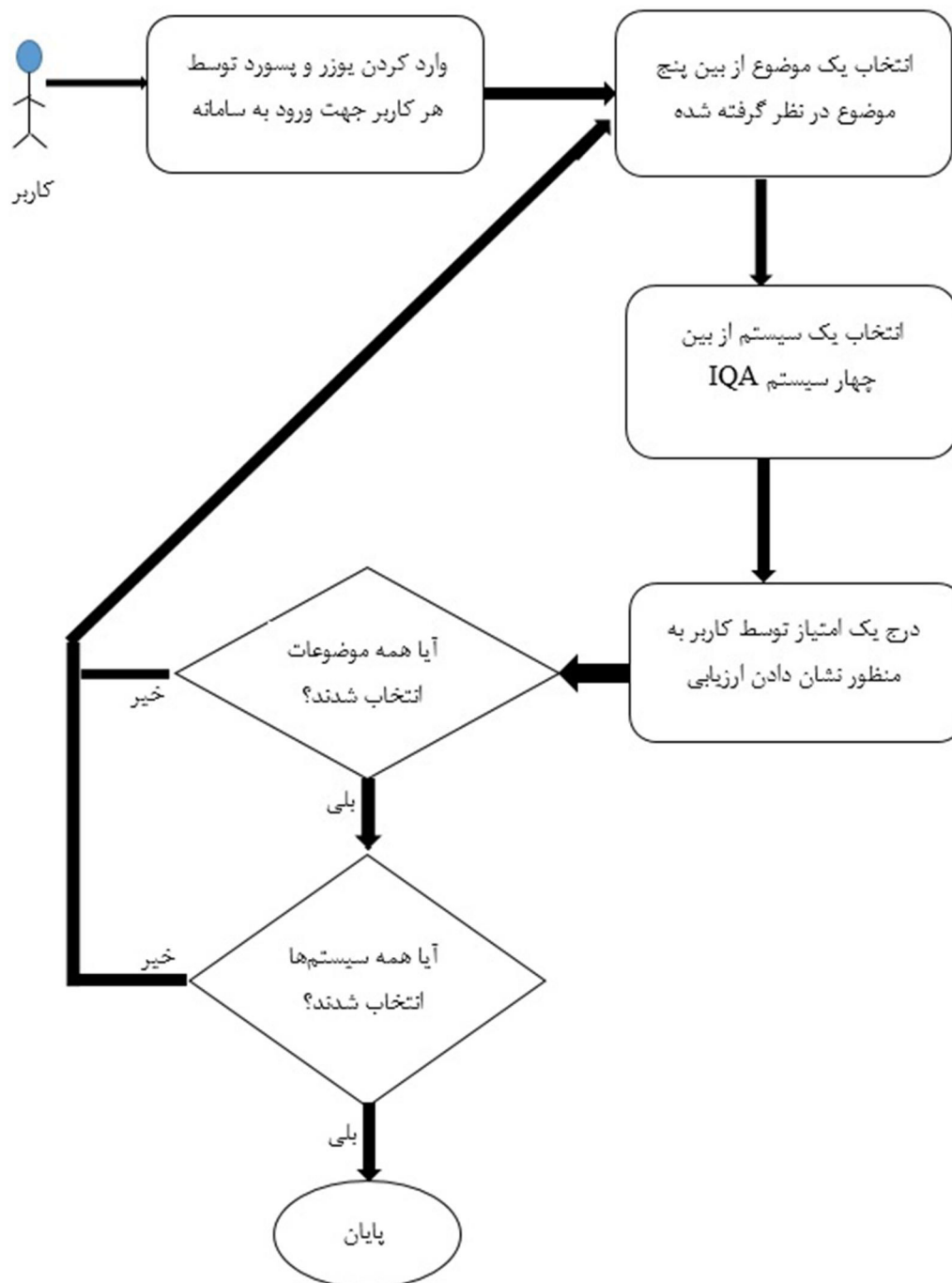
## فصل چهارم - تجزیه و تحلیل یافته‌های تحقیق

#### ۴-۱ مجموعه داده:

به دلیل فقدان مجموعه داده استاندارد در زمینه ارزیابی سیستم های IQA، نیاز به ایجاد یکسری متن خروجی از این سیستم ها و تشکیل یک مجموعه داده از سؤالات رد و بدل شده بین سیستم و کاربر با برچسب گذاری مناسب (تخصیص امتیاز کاربر به سیستم) بود. بر این اساس، علاوه بر سیستم تعاملی پایه طراحی شده در آزمایشگاه تکنولوژی زبان های طبیعی<sup>۱</sup> دانشگاه شاهرود، از سه سیستم تعاملی دیگر در ایجاد پایگاه داده مورد استفاده قرار گرفتند. برای یکپارچه سازی شرایط کار با این سیستم ها و راحتی کاربران، سامانه ای تحت وب طراحی گردید که متن تبادل شده و امتیاز داده شده توسط کاربران به سیستم ها را به صورت خودکار در پایگاه داده ای ذخیره نماید. شکل ۴-۱ چارچوب فرآیند استفاده از این سامانه طراحی شده را نمایش می دهد. در این راستا تعداد ۱۲۰ کاربر برای ۵ موضوع مختلف از سیستم ها استفاده نموده و مکالمات ردوبدل شده (با توجه به موضوع مکالمات) هر یک به صورت جداگانه در پایگاه داده ذخیره گردید. از این مجموعه ۶۰۰ تایی، ۵۴۰ نمونه توسط فرد خبره به عنوان نمونه مناسب تر انتخاب شد. شرکت کنندگان در این دوره ارزیابی شامل دانشجویان دانشگاه و زبان آموزان یک موسسه بودند. با توجه به اینکه یک سیستم با زبان فارسی و سه سیستم دیگر با زبان انگلیسی عمل می نمود، از افرادی در این راستا بهره گرفته شد که در حد متوسط، به زبان انگلیسی و استفاده از کامپیوتر مسلط باشند. یک دوره کوتاه آموزشی برای کاربران در مورد شرایط کار با سیستم ها و نحوه پاسخ دهی به سؤالات نیز برگزار گردید. تعداد ۹۰ نفر از شرکت کنندگان دارای مدرک لیسانس، ۲۷ نفر فوق لیسانس و ۳ نفر دانشجوی دکتری از رشته های مختلف بودند. همچنین جهت ارزیابی مناسب تر، یک پایگاه داده از مکالمات انگلیسی روزمره با موضوعات مختلف شکل گرفت. تعداد مکالمات در نظر گرفته شده ۲۴۵۰ مکالمه با ۵ موضوع مختلف بود. تمامی مکالمات دارای انسجام متنی بودند تا مکالمات واقعی به عنوان مکالمات مناسب در نظر گرفته شود. برای تشکیل مکالمات نامناسب در پایگاه داده (مکالماتی که از نظر کاربر امتیاز پایینی دارند) به صورت تصادفی

<sup>۱</sup>www.Hltech.ir

بعضی از خطوط مکالمات و کلمه‌های دیگر آن با دیگر مکالمات جابجا گردید و به عنوان مکالمه نامناسب در پایگاه داده ذخیره گردید تا روابط حاصل نیز بر روی این پایگاه داده مورد آزمون قرار گیرد.



شکل ۴-۱ نمایشی از فرآیند استفاده از صفحات مختلف سامانه تحت وب طراحی شده برای ارزیابی

جدول ۱-۴ یک نمونه تولید شده از یک مکالمه در سیستم IQA

کاربر	سیستم
اگر معدل دانشجوی کمتر از ۱۴ شود چه باید کرد؟	دانشجو مربوط به چه مقطعی است؟
دانشجو کارشناسی ارشد	منظور شما معدل کل هست؟
خیر، معدل یک‌ترم	در صورت کمتر شدن معدل از ۱۴ در هر نیمسال دانشجو مشروط تلقی می‌گردد.
دانشجوی کارشناسی چطور؟	در صورت کمتر شدن معدل از ۱۲ در هر نیمسال دانشجو مشروط تلقی می‌گردد.
چندترم متوالی مشروط شدن دانشجو اخراج می‌شود؟	دانشجو کارشناسی یا ارشد؟
دانشجو کارشناسی	در صورتی که دانشجو کارشناسی سه ترم متوالی یا چهار ترم غیر متوالی مشروط شود در آستانه اخراج قرار می‌گیرد.

جدول ۱-۴، یک نمونه از خروجی متن ذخیره شده (بدون فرمت) از تعامل با سیستم پایه را نشان می‌دهد. جدول ۲-۴ و ۳-۴ به ترتیب یک نمونه از مکالمه تولید شده با سیستم BOBdialogue و یک نمونه از مکالمات روزمره با امتیاز خوب و بد را نمایش می‌دهد. همچنین جدول ۴-۴ یک نمونه از داده استفاده شده در آموزش سیستم تعاملی پایه را نمایش می‌دهد.

جدول ۲-۴ یک نمونه از مکالمه تولید شده در سیستم Bobdialogue

B: Hello, I am BoB. I can tell you how to use the Library's services e.g., how to search the OPAC, borrow a book or where to find things in the Library. Please ask me questions in English - you will help me to learn!

2 U: where can I find books about morphology?

3 B: The OPAC offers two possibilities to search by topics. Both can be found in the advanced search menu. Searching for subject is the first possibility. And the second one ...

4 U: how do I use opac

5 B: First of all: You can access the OPAC wherever you have internet access. More details depend on what you are looking for ...

6 U: i like to find and borrow a specific book

7 B: Could you rephrase your question, please.

8 U: hod do I find and borrow a specific book

9 B: If you want to borrow books, DVDs or journals you need to take them form the shelf yourself and check them out at the information desk. If you want them to be sent to Brixen-Bressanone, Bozen-Bolzano, Bruneck-Brunico or Schlanders-Silandro you can [order] them via the OPAC.

10 ...

11 U: how will I know when the books I ordered are ready to be borrowed

12 B: If you stated while ordering, that the book should be reserved for you, we will contact you as soon

<p>as it arrives.</p> <p>13 U: How will you contact me?</p> <p>14 B: Gosh, I didn't quite get that. Strange, since I am actually pretty intelligent. Well, I will ponder upon that further and do some research, so the issue will make more sense to me on your next visit.</p> <p>15 U: when are the libraries open</p> <p>16 B: You can find our current opening hours on [this] website.</p> <p>17 U: are there special holidays when the libraries are closed</p> <p>18 B: On holidays my colleagues are allowed to stay at home. I'm of course always here. As well as our online offer. You can find our current opening hours on [this] website.</p> <p>19 ...</p> <p>20 U: for how long can I borrow library material</p> <p>.</p> <p>.</p> <p>.</p> <p>.</p>
--

جدول ۳-۴ یک نمونه از مکالمه مناسب و نامناسب

مکالمه با امتیاز خوب	مکالمه با امتیاز بد
<p>&lt;conversation id="2" subject="Viewing Houses with a Realtor"&gt;</p> <p>&lt;personA&gt;I have a good feeling about this house.&lt;/personA&gt;</p> <p>&lt;personB&gt;Yes, I liked it the first moment I pulled up to it.&lt;/personB&gt;</p> <p>&lt;personA&gt;Then you're not going to go to sleep.&lt;/personA&gt;</p> <p>&lt;personB&gt;If you like the outside, you are going to really love the inside.&lt;/personB&gt;</p> <p>&lt;personA&gt;What a beautiful home!&lt;/personA&gt;</p> <p>&lt;personB&gt;You'll notice that the window treatments, carpeting, and drapes are all new.&lt;/personB&gt;</p> <p>&lt;personA&gt;I like the way the blinds give you privacy from the street.&lt;/personA&gt;</p> <p>&lt;personB&gt; Follow me into the kitchen. You will love it.&lt;/personB&gt;</p> <p>&lt;personA&gt;I love that they put a wine storage area in the kitchen.&lt;/personA&gt;</p> <p>&lt;personB&gt;The best part is the bedroom and attached bathroom.&lt;/personB&gt;</p> <p>&lt;personA&gt;I love the relaxing colors in the tile and floor covering!&lt;/personA&gt;</p>	<p>&lt;conversation id="140" subject="Asking about Public Transportation"&gt;</p> <p>&lt;personA&gt;Can you give me a little more information about your apartment?&lt;/personA&gt;</p> <p>&lt;personB&gt;Weren't you taught that yellow means slow down, not speed up?&lt;/personB&gt;</p> <p>&lt;personA&gt;What kind of public transportation is near your apartment?&lt;/personA&gt;</p> <p>&lt;personB&gt;You should take a break.&lt;/personB&gt;</p> <p>&lt;personA&gt;Can I get you anything to drink?&lt;/personA&gt;</p> <p>&lt;personB&gt;I'm sorry. If you check online, you can get that kind of information.&lt;/personB&gt;</p> <p>&lt;personA&gt;Okay, I'll just go online.&lt;/personA&gt;</p> <p>&lt;personB&gt;I don't know anything for certain. We always do a double check if there is a question.&lt;/personB&gt;</p> <p>&lt;Score&gt;Bad&lt;/score&gt;</p>

جدول ۴-۴ یک نمونه از داده آموزشی استفاده شده در آموزش سیستم تعاملی پایه

<p>برگزاری دوره کارشناسی ارشد الکترونیکی به شیوه آموزش محور می‌باشد.</p> <p>موسسه می‌تواند در صورت وجود زیرساخت‌های لازم حداکثر تا ۲۰٪ دانشجویان برتر خود را بر اساس عملکرد آموزشی آنان به میزان ظرفیت اساتید و با تشخیص و تایید گروه آموزشی ذیربط و به شرط گذراندن حداقل دو نیم‌سال تحصیلی معادل ۱۲ واحد از دروس غیرجبرانی در قالب شیوه آموزشی پژوهشی بپذیرد.</p> <p>دانش‌آموختگان این دوره‌ها می‌توانند در مقاطع بالاتر ادامه تحصیل دهند و از مزایای استخدامی و اجتماعی همانند سایر دانش‌آموختگان برخوردار گردند درج عنوان نوع دوره آموزش الکترونیکی در مدارک تحصیلی دانش‌آموختگان الزامی است.</p> <p>صدور معافیت تحصیلی برای دانشجویان نظام یادگیری الکترونیکی مشروط به شرکت در آزمون برگزار شده توسط سازمان سنجش و مصوب بودن مقطع رشته و موسسات توسط شورای گسترش آموزش عالی بلامانع می‌باشد.</p> <p>آموزش در دوره‌های الکترونیکی در تمام دانشگاه‌های کشور مبتنی بر نظام واحدی است.</p> <p>هر واحد درسی در یادگیری الکترونیکی به ازای ۱۷-۱۵ ساعت آموزشی حداقل ۵ ساعت تولید محتوای مفید علمی به صورت الکترونیکی است که به همراه مثال‌ها تمرین‌ها تعاملات علمی و ارزیابی پایان هر جلسه ۱۷-۱۵ ساعت آموزشی را تشکیل می‌دهد.</p>
--

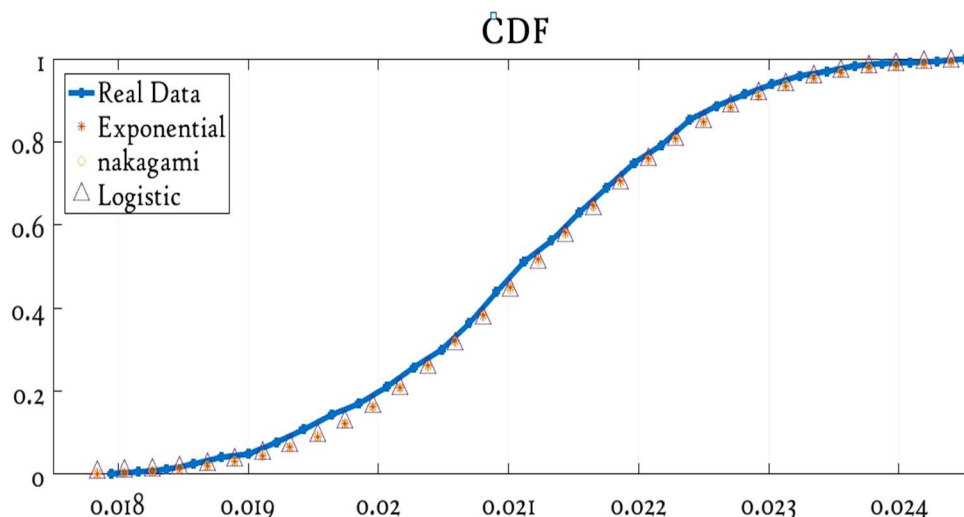
#### ۴-۲ آزمایش شماره ۱: نتایج حاصل از تعیین انسجام متن در سیستم‌های IQA

همانطور که در فصل قبل اشاره شد، نظر به اینکه در الگوریتم محاسبه انسجام، بیش از یک نمونه از مکالمه‌های منسجم و غیرمنسجم در اختیار داریم. مجموعه ویژگی‌های به دست آمده را به دو گروه آموزش و آزمون تقسیم و بهترین تابع چگالی احتمال که بر داده‌ها منطبق می‌باشد، تعیین گردید. برای اینکار، ۱۸ تابع چگالی احتمال مختلف مانند لاپلاس، گاما، گوسی و توابع دیگر در نظر گرفته شد. و بر روی تک‌تک ویژگی‌ها تست گردید تا منطبق‌ترین تابع چگالی بدست آید. از بین ۱۸ تابع در نظر گرفته شده و با توجه به فرم هیستوگرام به دست آمده، سه تابع چگالی احتمال نمایی<sup>۱</sup>، ناکاگامی<sup>۲</sup> و لجستیک<sup>۳</sup> به عنوان کاندیداهای احتمالی انتخاب شدند. یک نمونه تابع توزیع تجمعی برای هیستوگرام به دست آمده و فرم تقریب زده شده این سه تابع برای یک ویژگی از یک مکالمه در شکل

<sup>۱</sup> Exponential

<sup>۲</sup> Nakagami

<sup>۳</sup> Logistic



شکل ۴-۲. تابع توزیع تجمعی ۳۲ نقطه‌ای ویژگی اول برای یک مکالمه و سه تابع توزیع تجمعی تقریب زده شده

۴-۲ آورده شده است. همان‌طور که در این شکل نشان داده شده، هر سه تابع انطباق بسیار عالی با فرم اطلاعات واقعی دارند. برای اینکه بررسی شود کدام یک از این سه تابع چگالی احتمال انطباق بهتری با اطلاعات موجود دارند چند معیار سخت‌گیرانه آماری به عنوان سنجش میزان انطباق در نظر گرفته شد. در مرحله اول برای ارزیابی این سه تابع بدست آمده، میزان نیکویی برازش<sup>۱</sup> بین منحنی اطلاعات واقعی و منحنی توزیع تجمعی سه تابع تقریب زده شد. این معیار به ازای تمام ویژگی‌ها و تمامی مکالمات و برای دو دسته اطلاعات آموزش و آزمون به صورت مجزا محاسبه شد. علاوه بر آن در مواردی که اطلاعات هنگام تطبیق با تابع چگالی احتمال همخوانی نداشته باشد و نتواند تابع مناسبی برای پوشش اطلاعات واقعی با یک تابع چگالی احتمال بخصوص یافت نماید یک پیغام خطا مبنی بر عدم تطابق اطلاعات با تابع مفروض در برنامه نوشته شده در نظر گرفته شد. نتایج جدول ۴-۵ برای سه تابع در نظر گرفته شده، نشان می‌دهد که تقریباً هیچ خطای تطبیقی در سه حالت رخ نداده (در بدترین حالت ۱۱۸ خطا در مجموع ۱۸۷۲۰ حالت داریم که زیر ۱٪ محسوب می‌شود و قطعاً می‌توان هر سه تابع چگالی احتمال را در این وضعیت بسیار دقیق ارزیابی کرد) و معیار نیکویی برازش نیز به طرز فوق‌العاده مطلوبی برای هر سه تابع بالاست.

<sup>۱</sup> Goodness of fit

جدول ۴-۵ نتایج تطبیق سه تابع چگالی احتمال با کل ویژگی‌ها در سه گروه مکالمه انتخاب شده

تابع چگالی احتمال	مجموع مکالمه‌ها	تعداد ویژگی	خطای تطبیق	نیکویی برازش
Exponential	۳۹۰	۴۸	۱۱۸	۰/۹۰۷
Nakagami	۳۹۰	۴۸	۲۱	۰/۹۴۷
Logistic	۳۹۰	۴۸	۸۹	۰/۹۱۲

با توجه به اعداد بسیار مناسب نیکویی برازش نمی‌توان تفاوت معنی‌داری بین سه تابع یافت و به همین دلیل نیاز به آزمون‌های آماری دقیق‌تر وجود دارد. قبل از بررسی دیگر آزمون‌های آماری نیاز بود تا پایداری توابع بدست آمده بررسی گردد. بنابراین در اولین گام با استفاده از توابع چگالی احتمال به دست آمده، اطلاعات تصادفی جدید تولید گردید (لازم به ذکر است که ما طول این بردار تصادفی را به صورت پیش فرض ۲۰۰ در نظر گرفتیم). سپس بین اطلاعات تولید شده با تابع چگالی احتمال به دست آمده از نمونه‌های آموزش یک کلاس از مکالمه و نمونه‌های آزمایش واقعی همان کلاس از مکالمه، به وسیله آزمون‌های آماری، یکسان بودن تابع احتمال را کنترل می‌کنیم. بدیهی است باید جواب مثبت باشد و آزمون متناظر یکسان بودن دو نمونه را اعلام کند. در غیر این صورت بدان معنی است که تابع چگالی احتمال تقریب زده شده برای نمونه‌های آموزش فرم مناسبی نداشته و پایدار نیست. نتایج حاصل از این آزمایش در جدول ۴-۶ آورده شده است.

جدول ۴-۶ بررسی تطبیق یک کلاس از مکالمه با خودش در سه تابع چگالی احتمال مورد بررسی

تابع چگالی احتمال	مجموع مکالمات	تعداد ویژگی	تعداد تطبیق	خطای تطبیق	درصد خطای تطبیق
Exponential	۳۹۰	۴۸	۱۸۷۲۰	۴۵۱	٪۲.۵
Nakagami	۳۹۰	۴۸	۱۸۷۲۰	۲۱۲	٪۱.۱
Logistic	۳۹۰	۴۸	۱۸۷۲۰	۳۷۴	٪۲



در مرحله بعدی، سه آزمون آماری شامل کولموگروف-اسمیرنف<sup>۱</sup>، اندرسون دارلینگ<sup>۲</sup> و کرامر وان میس<sup>۳</sup> استفاده شد که هر سه از معیارهای متداول بررسی یکسان بودن توزیع بین دو دسته اطلاعات با طول یکسان محسوب می‌شوند. بر اساس نتایج به دست آمده می‌توان مشاهده نمود که هر سه تابع با دقت بسیار زیادی توانسته‌اند تابع چگالی احتمال یک مکالمه را پیش‌بینی و مشخص کنند (جدول ۴-۷). در این مرحله به سراغ ایجاد تمایز بین یک کلاس از مکالمه و سایر کلاس‌ها می‌رویم. تابع چگالی احتمال مناسب باید بین مکالمه یک کلاس و کلیه کلاس‌های دیگر در مکالمه، تفاوت قائل شده و هر سه آزمون یاد شده شباهت اطلاعات در دو مکالمه مختلف را رد کنند. طبعاً تعداد مقایسه‌های صورت گرفته، در این حالت بسیار بالاتر است چرا که یک کلاس از مکالمه با تمامی کلاس‌های دیگر در تمام ویژگی‌ها باید مقایسه و امتیاز دهی شود. از آنجا که درصد انطباق هر سه تابع بسیار بالا بود، در این مرحله به سراغ درصد درستی کل نرفتیم و تک‌تک اشتباهات در هر سه تابع را به صورت جداگانه شمارش کرده‌ایم. در این مرحله مجموعاً ۱۳۰ مکالمه برای هر کلاس داشتیم که هر کلاس از یک مکالمه با مکالمه هم گروه خود در سه تابع چگالی احتمال و به ازای هر تابع دو بار در ۴۸ ویژگی مقایسه شد. مجموعاً در هر حالت ۱۸۷۲۰ مقایسه صورت می‌پذیرد.

جدول ۴-۷ مقایسه بین تابع چگالی احتمال کلاس‌های مکالمه مختلف و آزمون‌های آماری

تابع چگالی احتمال	کولموگروف-اسمیرنف		کرامر وان میس		اندرسون-دارلینگ		تعداد تطبیق
	درصد درستی	اشتباه	درصد درستی	اشتباه	درصد درستی	اشتباه	
Exponential	۹۶/۰۳	۷۴۲	۹۶/۸۸	۵۸۴	۹۷/۴۸	۴۷۳	۱۸۷۲۰
Nakagami	۹۶/۵	۶۵۴	۹۸/۳	۳۱۲	۹۸/۸۳	۲۱۹	۱۸۷۲۰
Logistic	۹۵/۰۱	۹۱۷	۹۶/۶	۶۳۱	۹۷/۴۰	۴۸۶	۱۸۷۲۰

<sup>1</sup> Kolmogorov-Smirnov

<sup>2</sup> Anderson-Darling

<sup>3</sup> Kramer Van-Miss

نتیجه به دست آمده باز هم فوق‌العاده عالی است. اما در این حالت بین سه تابع می‌توان تفاوت‌های مشخصی دید و تابع مناسب‌تر را انتخاب نمود. نتیجه حاصل از این کار در جدول ۴-۷ نمایش داده شده است. بر اساس نتایج این جدول تابع چگالی احتمال ناکاگامی با مجموع ۱۱۸۵ اشتباه نسبت به دو تابع دیگر که هر دو بیش از ۱۶۰۰ اشتباه داشته‌اند بهتر عمل کرده است. بنابراین ویژگی‌های سیزدهم تا چهاردهم نشان دادند که می‌توانند در تعیین انسجام یک مکالمه تولید شده از سیستم پرسش و پاسخ تعاملی مورد استفاده قرار گیرند. در حقیقت این آزمایش اثباتی بر در نظر گرفتن معیار انسجام با توجه به ویژگی‌های پیشنهادی می‌باشد. این روش یکی از نوآوری‌های انجام شده در این رساله است که در کارهای قبلی مشاهده نگردیده بود.

#### ۴-۳- آزمایش شماره ۲: یافتن بهترین مدل بر اساس ویژگی‌های پیشنهادی به

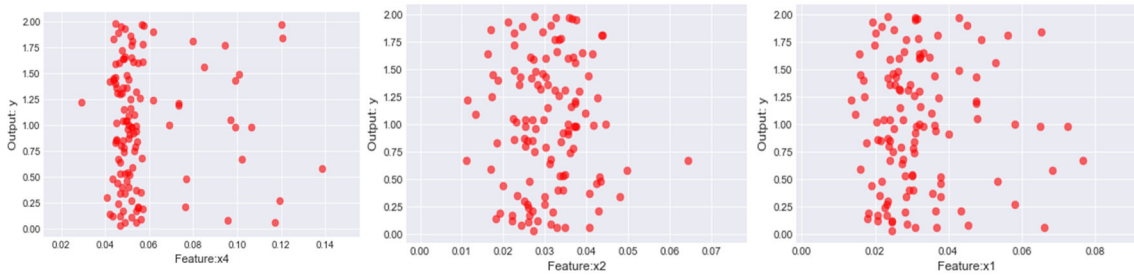
##### کمک رگرسیون

مرحله ساخت ویژگی‌ها و مدل‌سازی به صورت همزمان انجام شده است (یعنی کل ویژگی‌های معرفی شده در فصل سوم در یک مرحله حاصل نگردید بلکه به مرور زمان به این ویژگی‌ها اضافه گردید). بنابراین قبل از رسیدن به نوع مدل نیاز است تا به مقادیر بدست آمده برای هر ویژگی توجه نموده تا در انتخاب نوع مدل با دقت بیشتری صورت پذیرد. مقادیر مربوط به برخی از ویژگی‌ها در به دو صورت نمودار پراکندگی و میله‌ای هم برای حالتی که ویژگی‌ها دسته‌بندی نشده‌اند و هم برای حالتی که دسته‌بندی شده‌اند در شکل ۴-۳ نمایش داده شده است. حال در ادامه به بررسی نتایج حاصل از بعضی مدل‌های رگرسیونی می‌پردازیم.

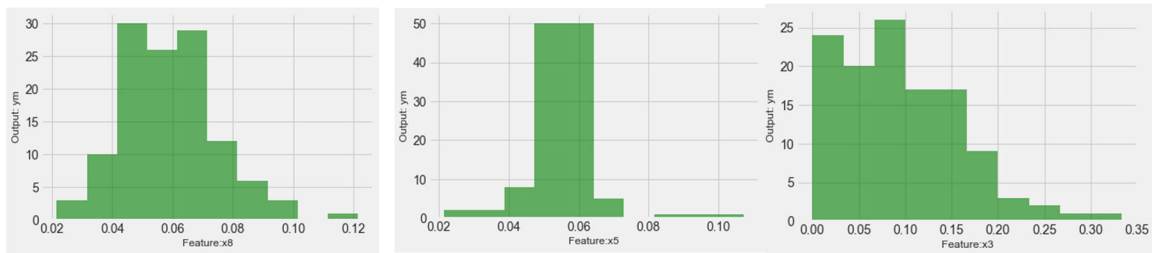
#### ۴-۳-۱ رگرسیون خطی برای پیش‌بینی نظرات

رگرسیون خطی به بررسی رابطه یک متغیر مستقل (پیش‌بین) و یک متغیر وابسته می‌پردازد. این نوع از رگرسیون یکی از تکنیک‌های مدل‌سازی شناخته شده است که بطور گسترده در مسائل مختلف

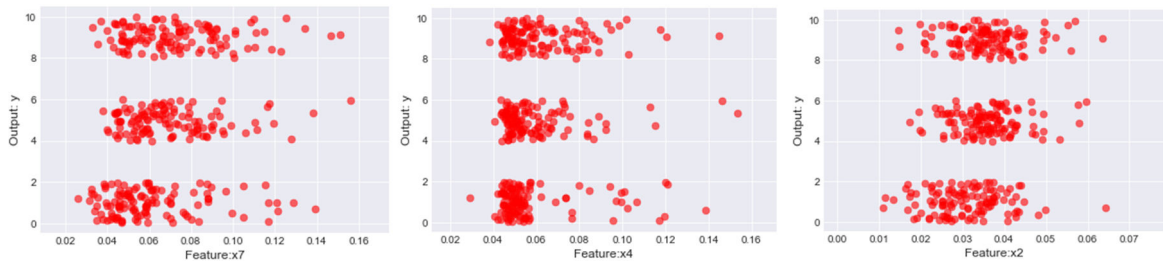
به کار گرفته می‌شود. همچنین رگرسیون خطی، معمولاً در میان چند موضوع اولی است که افراد هنگام پیش‌بینی و مدل‌سازی از آن استفاده می‌کنند. رگرسیون خطی به دو صورت رگرسیون خطی ساده و رگرسیون خطی چندگانه (چند متغیره) مطرح می‌گردد که نوع آن باید متناسب با مسئله انتخاب شود.



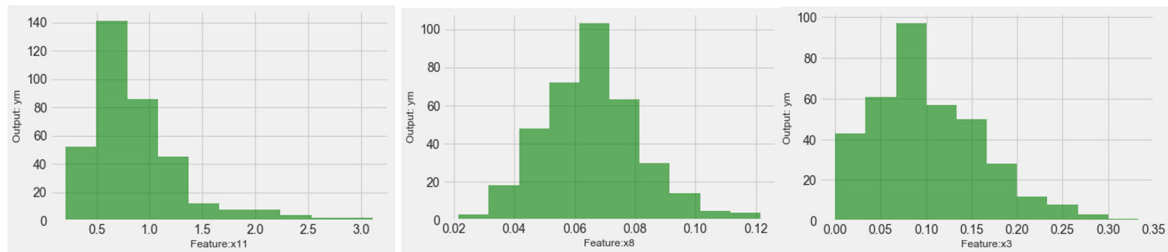
الف) نمودار پراکندگی مربوط به مقادیر ویژگی‌های اول، دوم و چهارم برای کلاس مکالمات ضعیف



ب) نمودار میله‌ای مربوط به مقادیر ویژگی‌های سوم، پنجم و هشتم برای کلاس مکالمات ضعیف



ج) نمودار پراکندگی مربوط به مقادیر ویژگی‌های دوم، چهارم و هفتم برای مکالمات



د) نمودار میله‌ای مربوط به مقادیر ویژگی‌های سوم، هشتم و یازدهم برای مکالمات

شکل ۴-۳ نمودار میله‌ای و پراکندگی برخی از ویژگی‌ها

برای رسیدن به بهترین خط، روش‌های متعددی وجود دارد. از جمله روش‌ها می‌توان به روش همزمان<sup>۱</sup>، گام به گام<sup>۲</sup>، حذف<sup>۳</sup> و روش مربعات خطا<sup>۴</sup> که در آن هدف به حداقل رساندن مجموع مربعات انحراف عمودی از هر نقطه داده به خط تا کمترین مقدار می‌باشد، اشاره نمود. چند نکته قابل توجه در این حوزه وجود دارد که در زمان استفاده از رگرسیون خطی باید به آن توجه نمود.

الف- بین متغیرهای مستقل و وابسته رابطه خطی وجود دارد.

ب- رگرسیون چندگانه از هم خطی بین متغیرهای مستقل<sup>۵</sup>، خود همبستگی<sup>۶</sup>، ناهمسانی واریانس<sup>۷</sup> رنج می‌برد.

ج- رگرسیون خطی به نقاط دور افتاده بسیار حساس است و این می‌تواند به شدت بر خط رگرسیون و در نهایت ارزش پیش بینی شده تاثیر بگذارد.

د- هم خطی بین متغیرهای مستقل می‌تواند واریانس برآوردهای ضریب را افزایش دهد و تخمین‌ها را بسیار حساس به تغییرات جزئی در مدل نماید که در نتیجه، سبب ناپایداری برآورد ضرایب می‌شود.

اما قبل از انتخاب هر یک از این مدل‌ها چند سوال بوجود می‌آید که باید در موقع انتخاب رگرسیون‌ها به آن‌ها توجه نماییم. بطور مثال برای رگرسیون خطی چند جمله‌ای باید به سوالاتی مانند چگونگی تصمیم‌گیری در مورد اینکه چه چندجمله‌ای باید در نظر گرفته شود؟ چه زمانی باید ادامه دادن برای رسیدن به چند جمله‌ای مناسب متوقف شود اگر ما با ترکیب درجه اول، درجه دوم، شرایط درجه ۳ و درجه‌های بالاتر یک به یک شروع کنیم؟ و یا اینکه چگونگی تصمیم‌گیری در مورد اینکه آیا هر یک از ترم‌های متقابل، آیا مهم هستند یا خیر؟ (یعنی ما فقط نیاز به ترم‌های  $X_1^2$ ،  $X_2^3$  یا  $X_1 \cdot X_2$  و  $X_1^2 \cdot X_3$  را داریم یا به همه ترم‌های متقابل نیازمندیم) و در نهایت، آیا ما باید به صورت دستی معادلات توابع را

<sup>1</sup> Enter Method

<sup>2</sup> Stepwise Method

<sup>3</sup> Remove Method

<sup>4</sup> Square Method

<sup>5</sup> multicollinearity

<sup>6</sup> autocorrelation

<sup>7</sup> heteroskedasticity

برای تمام این تغییرات چندجمله‌ای بنویسیم و آن‌ها را به مجموعه داده اضافه کنیم. بنابراین اهمیت تنظیمات را نمی‌توان در نظر نگرفت چون که تنظیم پارامترها، مفهوم اصلی یادگیری ماشین است.

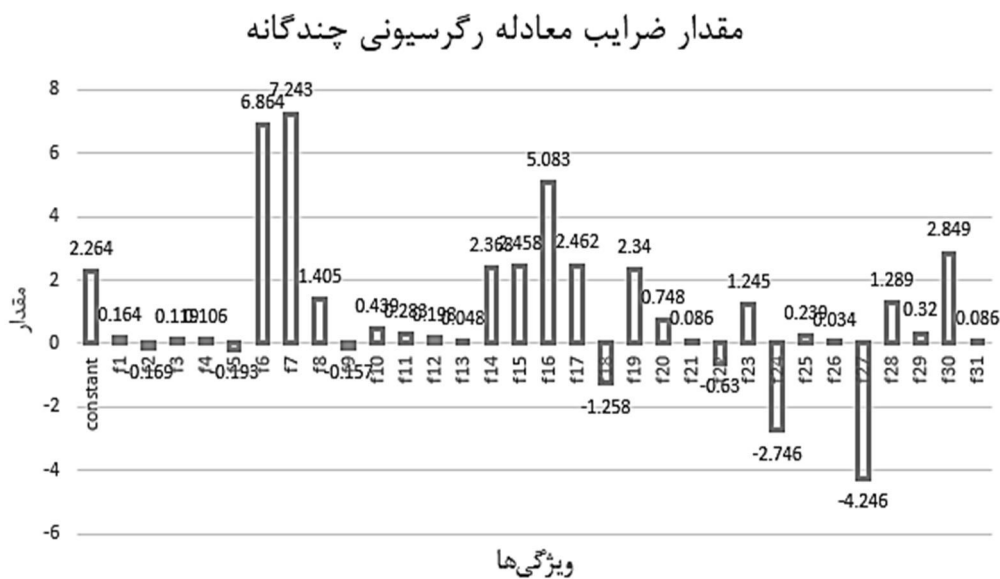
#### ۴-۳-۱-۱ رگرسیون خطی چندگانه

در رگرسیون خطی اگر تعداد متغیرهای مستقل در رابطه بیشتر از یک شود، مدل رگرسیون، خطی چندگانه نامیده می‌شود. معادله رگرسیون خطی ساده به شکل  $Y = AX + B$  و رگرسیون خطی چندگانه به صورت  $Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$  است که در آن  $X$ ها ویژگی‌های استخراج شده و  $b$  ضرایب تأثیر هر یک از ویژگی‌ها می‌باشد. در نتیجه این ضرایب می‌بایست به گونه‌ای محاسبه گردند که معیار حداقل مربعات خطا را تأمین نمایند. مقدار حاصل برای  $a$  بیانگر مقادیر پیش‌بینی شده  $Y$  با ثابت ماندن مقادیر  $X$  است. از سوی دیگر، با مقایسه اندازه مقادیر ضرایب  $b$  با یکدیگر اولویت و میزان تأثیر هر یک از عوامل مشخص می‌شود. همچنین علامت ضرایب هم بر تغییرات متغیر وابسته تأثیرگذار می‌باشد. در رگرسیون خطی چندگانه پنج فرضیه نرمال بودن<sup>۱</sup>، کاهش واریانس<sup>۲</sup>، متغیرهای توضیحی ثابت، استقلال<sup>۳</sup> و مشخصه مدل صحیح وجود دارد که باید به آن‌ها در زمان استفاده این مدل توجه نمود. در روش رگرسیون خطی ویژگی‌های استخراج شده به عنوان متغیر مستقل و نظرات انسانی به عنوان متغیر وابسته در نظر گرفته شدند. ضرایب رگرسیونی خطی به دست آمده با توجه به معادله رگرسیون خطی چندگانه برای در شکل ۴-۴ نمایش داده شده است. مقدار ضریب تعیین  $R^2$  برای معادله برابر ۰/۲۵، RMSE برابر ۰/۳۴ و MAPE برابر ۰/۱۵ به دست آمده است. همان‌طور که در شکل ۴-۴ نشان داده شده، مقادیر به دست آمده برای ضرایب معادله رگرسیونی، دارای علامت و مقدار متفاوتی برای هر متغیر می‌باشند.

<sup>1</sup> Normality

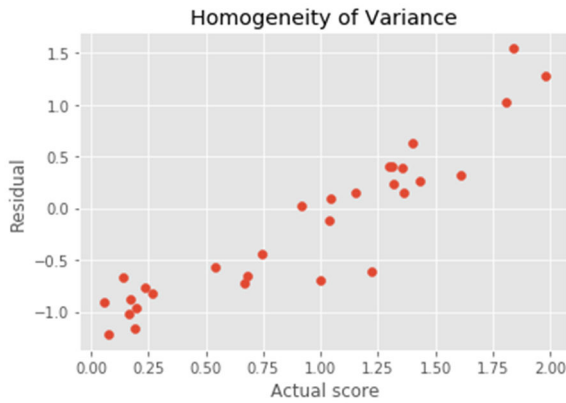
<sup>2</sup> Homogeneity of Variance

<sup>3</sup> Independence

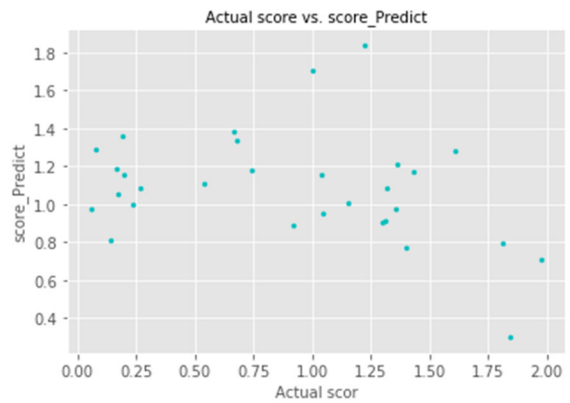


شکل ۴-۴ ضرایب معادله رگرسیون خطی چندگانه

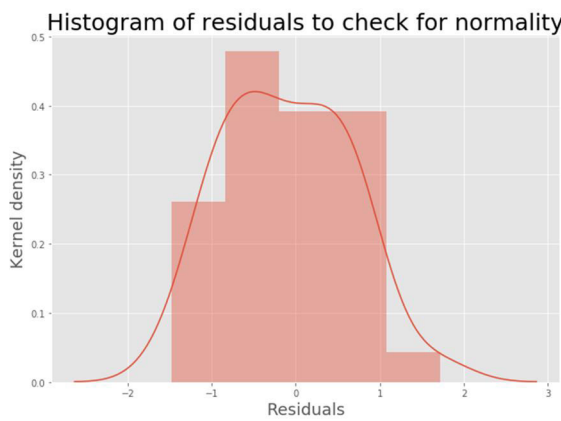
ویژگی‌هایی که دارای مقدار کم می‌باشند دارای تأثیر کمتر در نظرات انسانی و آن‌هایی که دارای مقادیر بزرگ‌تر هستند حاکی از تأثیر بیشتر این ویژگی‌ها در خروجی می‌باشند. همچنین با توجه به ضرایب به دست آمده، می‌توان تغییرات نظرات انسانی را نسبت به متغیرهای مستقل نیز به دست آورد. با توجه به معیارهای ارزیابی، همانطور که از نتایج پیداست، اعداد محاسبه شده برای ارزیابی خط رگرسیونی، از مقادیر قابل قبولی برخوردار نیستند که بتوان از این معادله خط به عنوان پاسخ نهایی یاد نمود. همچنین برای اطمینان از عدم رخداد بیش‌برازش، روش تصدیق تقاطعی برابر ۱۰ دسته بهره گرفته شد که نتایج حاصل، دوباره اعداد قابل قبولی را گزارش نکرد. نتایج حاصل در شکل ۴-۵ نمایش داده شده است. لذا روش‌های دیگر رگرسیونی مورد استفاده قرار گرفت.



(ب) نمودار کاهش واریانس



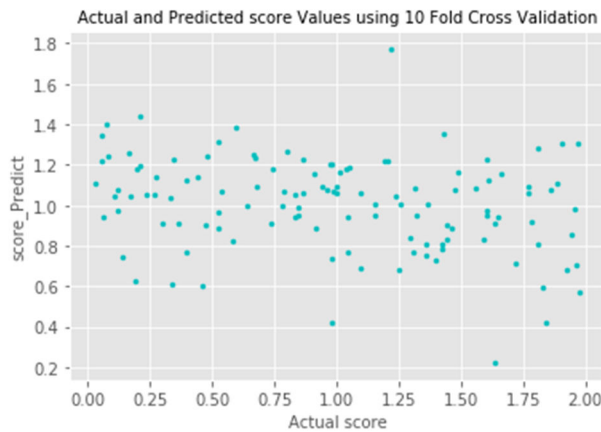
(الف) نمودار امتیاز واقعی در برابر امتیاز پیش‌بینی شده



(ت) هیستوگرام مربوط به باقیمانده‌ها جهت چک کردن نرمال بودن



(پ) نمودار ضرایب همبستگی بین ویژگی‌ها



(ج) نمودار امتیاز واقعی در برابر امتیاز پیش‌بینی شده حاصل از روش تصدیق تقاطعی

شکل ۴-۵ نتایج حاصل از رگرسیون خطی چندگانه

#### ۴-۳-۲ رگرسیون غیرخطی برای پیش‌بینی نظرات

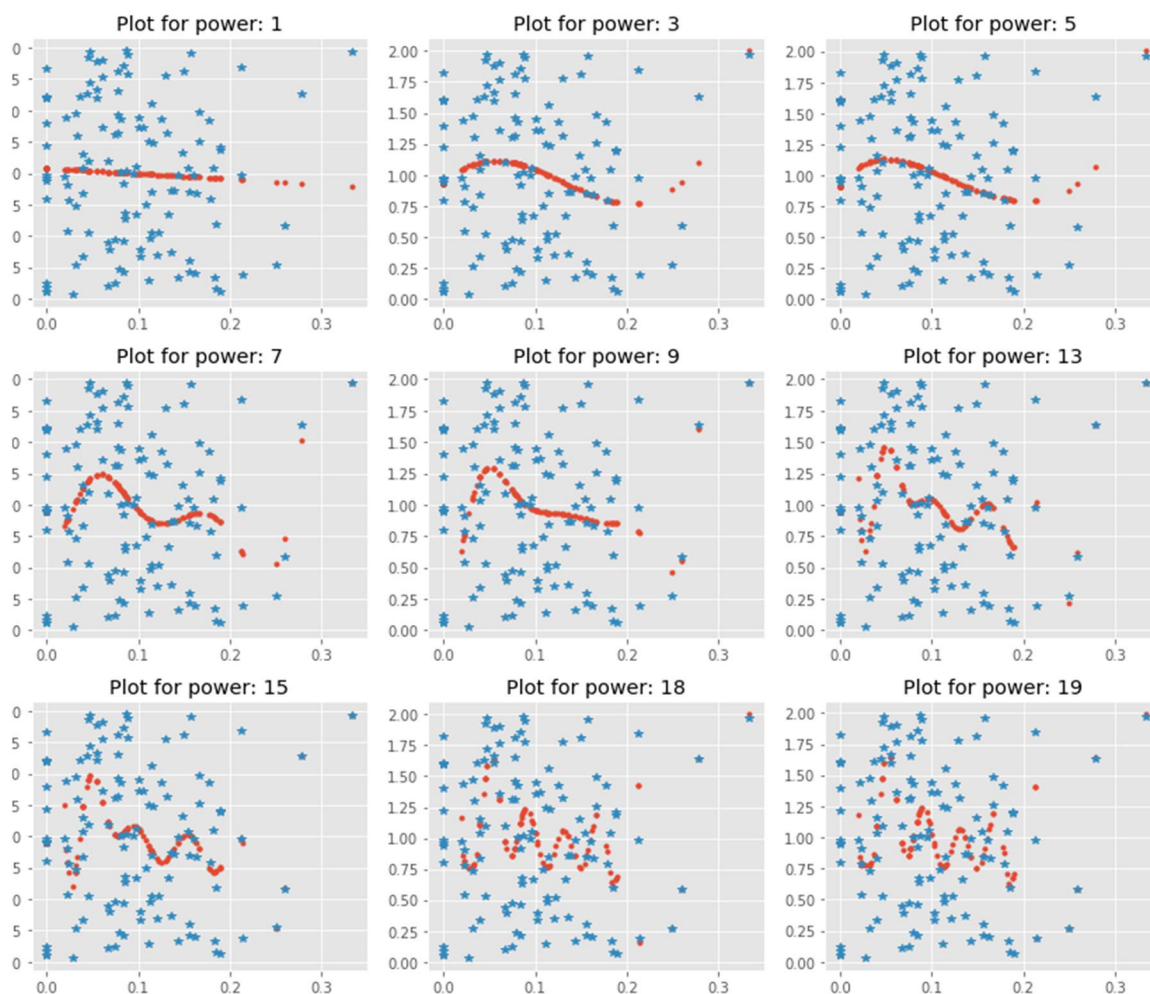
رگرسیون غیرخطی مدل‌های مختلفی دارد که از جمله آن می‌توان به مدل‌های درجه ۲ به بالا، چندجمله‌ای نمایی، توانی، لجستیک، لسو و... اشاره کرد که متناسب با مدل انتخاب شده معادلات آن‌ها متفاوت خواهد بود. متناسب با آزمایش‌ها، انواع مختلفی از معادلات رگرسیون غیرخطی برای انجام مدل‌سازی با مقادیر اولیه متفاوتی برای ضرایب در نظر گرفته شد و بر روی ماتریس ویژگی مورد آزمایش قرار گرفت تا بهترین مدل انتخاب گردد. در ادامه نتایج حاصل برخی از مدل‌های پیاده‌سازی شده آورده شده است.

#### ۴-۳-۱ رگرسیون چندجمله‌ای

اگر درجه توان یک متغیر در معادله رگرسیون بیشتر از یک در نظر گرفته شود، آن‌گاه، آن یک معادله رگرسیونی چندجمله‌ای<sup>۱</sup> خواهد بود. در این حالت، مناسب‌ترین خط، یک خط مستقیم نمی‌باشد بلکه یک منحنی خواهد بود که متناسب با نقاط داده‌ها است. نکته‌ای که باید در زمان استفاده از این معادله رگرسیونی به آن توجه داشت، وسوسه استفاده از چندجمله‌ای درجه بالاتر برای رسیدن به خطای کمتر است که این می‌تواند منجر به بیش‌برازش شود.

<sup>۱</sup> Polynomial





شکل ۴-۶ نتایج حاصل از انتخاب مدل با استفاده از رگرسیون چندجمله‌ای

بنابراین نیاز است که همیشه روابط را ترسیم کرده تا تناسب بین داده‌ها و منحنی را مشاهده و بررسی اطمینان از اینکه این منحنی با ماهیت مشکل متناسب است، باید دقت کافی را داشته باشیم. به خصوص برای منحنی‌ها، که باید به سمت انتهای آن‌ها نگاه کرده و مشاهده کرد که آیا این شکل‌ها و گرایش‌ها منطقی هستند یا خیر. بخشی از نتایج شبیه‌سازی در جدول و شکل ۴-۶ نمایش داده شده است. برای رسیدن به درجه مناسب برای معادله از رنج ۲ تا ۲۰ را تست نمودیم. مقدار ضریب تعیین  $R^2$  برای معادله برابر  $0/30$ ، RMSE برابر  $0/54$  و MAPE برابر  $9\%$  به دست آمده است. با این حال نتایج کسب شده حاکی از عدم انتخاب این مدل از رگرسیون بود.

جدول ۴-۶ نمایش برخی از ضرایب بدست آمده برای دستیابی به معادله خط رگرسیونی

	rss	intercept	coef_x_1	coef_x_2	coef_x_3	coef_x_4	coef_x_5	coef_x_6	coef_x_7	coef_x_8
model_pow_1	40	1	-0.43	NaN	NaN	NaN	NaN	NaN	NaN	NaN
model_pow_2	39	1.1	-2	6.7	NaN	NaN	NaN	NaN	NaN	NaN
model_pow_3	37	0.93	7.4	-84	2.10E+02	NaN	NaN	NaN	NaN	NaN
model_pow_4	37	0.92	8.1	-97	2.90E+02	1.20E+02	NaN	NaN	NaN	NaN
model_pow_5	37	0.91	11	1.60E+02	9.00E+02	2.50E+03	3.10E+03	NaN	NaN	NaN
model_pow_8	36	0.95	-47	3.20E+03	6.90E+04	7.30E+05	4.10E+06	1.20E+07	1.90E+07	1.10E+07
model_pow_9	35	0.96	1.20E+02	9.20E+03	2.60E+05	3.90E+06	3.40E+07	1.80E+08	5.60E+08	9.70E+08
model_pow_10	35	0.95	-36	8.40E+02	6.40E+04	2.80E+06	4.70E+07	4.30E+08	2.30E+09	7.00E+09
model_pow_11	35	0.95	29	6.60E+03	4.10E+05	1.10E+07	1.70E+08	1.60E+09	9.60E+09	3.60E+10
model_pow_12	35	0.92	2.40E+02	3.40E+04	1.80E+06	5.20E+07	9.10E+08	1.00E+10	7.60E+10	3.80E+11
model_pow_13	34	0.81	1.40E+03	1.90E+05	1.10E+07	3.60E+08	7.20E+09	9.60E+10	8.80E+11	5.60E+12
model_pow_14	34	0.75	1.90E+03	2.80E+05	1.70E+07	5.90E+08	1.30E+10	1.80E+11	1.90E+12	1.30E+13
model_pow_15	34	0.82	1.30E+03	1.70E+05	8.60E+06	2.20E+08	2.50E+09	8.20E+09	6.90E+11	1.10E+13
model_pow_16	33	1.1	2.00E+03	4.50E+05	4.30E+07	2.20E+09	7.40E+10	1.60E+12	2.60E+13	3.00E+14
model_pow_17	33	1.9	1.00E+04	2.10E+06	1.90E+08	1.00E+10	3.50E+11	8.50E+12	1.50E+14	1.90E+15
model_pow_18	31	0.48	1.50E+04	3.40E+06	3.60E+08	2.20E+10	8.90E+11	2.50E+13	5.30E+14	8.20E+15
model_pow_19	31	0.82	1.80E+04	4.30E+06	4.50E+08	2.80E+10	1.10E+12	3.30E+13	6.90E+14	1.10E+16

## ۴-۳-۲-۲ رگرسیون ریج

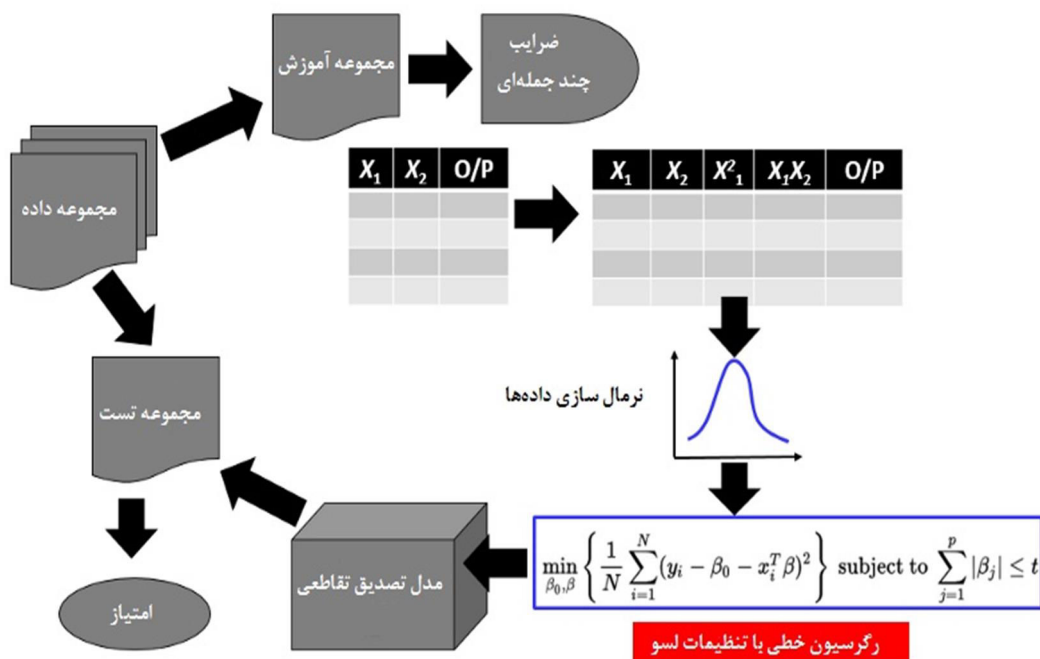
رگرسیون ریج روشی است برای برآورد ضرایب رگرسیون در مورد داده‌هایی که غیرمتعامد هستند و در مقایسه با برآوردهای کمترین مربعات به روش معمول، میانگین مربعات خطای کوچکتری دارند. برآوردهای کمترین مربعات ریج پایدار بوده، تحت تأثیر تغییرات جزئی در داده‌ها واقع نمی‌شوند. از آنجا که میانگین مربعات خطا در روش ریج کوچکتر است، مقادیر برآورد شده با این روش در مقایسه با روش معمول به مقادیر واقعی ضرایب رگرسیون نزدیکتر هستند. میزان برتری این روش در مقایسه با روش معمول در برآورد ضرایب، به مقادیر حقیقی این ضرایب در مدل بستگی دارد. از این روش نیز در جهت رسیدن به معادله مناسب استفاده گردید. برای مقایسه نتایج حاصل از این رگرسیون از چهار رگرسیون دیگر نیز استفاده گردید تا بتوان درک مناسب‌تری نسبت به نتایج حاصل شده داشت. مقادیر بدست آمده برای هر روش در جدول ۴-۷ آورده شده است. در این جدول روش‌های مختلف رگرسیونی توسط سه معیار با یکدیگر مقایسه شده‌اند که اعداد بدست آمده برای رگرسیون ریج نسبت به دیگر روش‌ها مناسب‌تر بوده و معادله خط بهتری را توصیف کرده است. با این وجود هنوز برای رسیدن به بهترین معادله خط مقداری فاصله وجود دارد. بنابراین می‌توان روش‌های دیگر را نیز مورد آزمایش قرار داد.

جدول ۴-۷ مقایسه نتایج رگرسیون ریج با چهار روش دیگر

	RMSE	R-squared	MAE
SGDRegressor	۰/۴۷	۰/۱۶۲	۲/۶۵
Ridge	۰/۲۴	۰/۶۷۶	۲/۰۷
ARDRegression	۰/۳۱	۰/۴۶۷	۲/۰۴
PassiveAggressiveRegressor	۰/۴۴	۰/۲۵	۲/۶۸
TheilSenRegressor	۰/۴۵۲	۰/۳۴۴	۲/۱۳

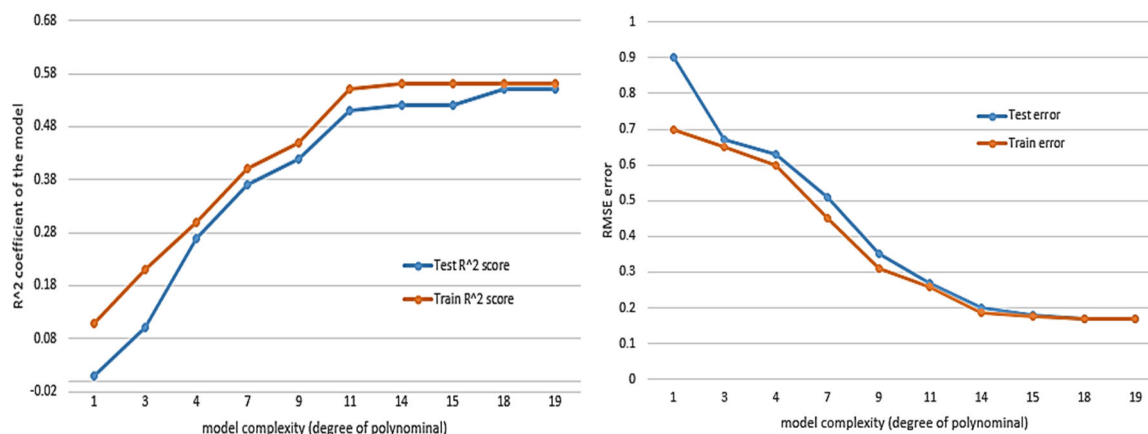
## ۳-۲-۳-۴ رگرسیون لسو

همانند رگرسیون ریج، رگرسیون لسو<sup>۱</sup> اندازه مطلق ضرایب رگرسیون را مورد بررسی قرار می‌دهد. علاوه بر این، قادر به کاهش تغییرپذیری و بهبود دقت مدل‌های رگرسیون نیز است. شکل ۴-۷ فرآیند استفاده از این رگرسیون را نمایش می‌دهد. رگرسیون لسو از رگرسیون ریج به گونه‌ای متفاوت است که از مقادیر مطلق در تابع جریمه به جای مربع استفاده می‌کند. این امر منجر به محدود کردن مجموع مقادیر مطلق برآوردها می‌شود که باعث می‌شود برخی از تخمین‌های پارامتری دقیقاً صفر شوند. پس از اعمال مجازات، برآوردها ای بیشتر به صفر مطلق تبدیل می‌شوند. این امر منجر به انتخاب متغیر از  $n$  متغیر می‌شود. که این یکی از مزیت‌های این روش از رگرسیون می‌باشد. بنابراین می‌توان از روش‌های کاهش ویژگی نیز استفاده نموده و سپس دوباره روش‌های رگرسیونی را تکرار کرد. همانطور که در این شکل نشان داده شده است ابتدا ماتریس ویژگی به دو مجموعه تست و آموزش به نسبت ۰/۷ تقسیم گردید. سپس ضرایب چندجمله‌ای برای مجموعه آموزش ایجاد گردید.



شکل ۴-۷ فرآیند استفاده از رگرسیون لسو

<sup>1</sup> Least Absolute Shrinkage and Selection Operator (lasso)



شکل ۴-۸ روند پیچیدگی مدل در برابر معیارهای ارزیابی برای مجموعه آموزش و تست با استفاده از رگرسیون لسو

در مرحله بعد داده‌ها نرمال‌سازی و با استفاده از رگرسیون لسو معادله خط برای حالت‌های مختلف محاسبه و توسط مجموعه تست مورد ارزیابی قرار گرفت. برای همه مدل‌ها، خطای آموزش (RMSE)، خطای تست و ضریب  $R^2$  برای معادلات حاصل گردید. مقایسه نتایج برای مدل‌های مختلف حاکی از این بود که با افزایش درجات چندجمله‌ای (پیچیدگی مدل) مقدار RMSE کمتر و مقدار ضریب  $R^2$  بیشتر شد که حاکی از بهتر شدن نتایج بود و اینکه این مدل رگرسیونی نتایج بهتری گزارش نموده است (شکل ۴-۸). همانطور که در دو نمودار مشاهده می‌کنید افزایش پیچیدگی مدل همراه با کاهش خطا برای مدل می‌باشد. بنابراین برای مدل‌سازی به درجات بالا از چندجمله‌ای نیاز داریم و مدل‌های خطی، درجه دوم یا حتی مکعبی برای برازش داده‌ها به اندازه کافی پیچیده نیستند. با این حال در درجات یازده به بالا تقریباً یک روند یکنواختی (تغییرات ناچیز) داریم که دیگر نیاز نیست به دلیل پیچیدگی بالا به سراغ درجات بالاتر برویم. از طرفی اگر به نمودار دقت نمایید با افزایش پیچیدگی مدل خطای تست به شدت بالا نمی‌رود. دلیل این امر در این واقعیت نهفته است که با استفاده از رگرسیون لسو، شرایط مرتبه بالاتر را حذف می‌کنیم. در حقیقت این مسئله یکی از مزایای کلیدی رگرسیون لسو است که برخی از ضرایب مدل را به جای اینکه فقط آن‌ها را کوچک نماید به طور دقیق صفر می‌نماید. بنابراین یک نوع انتخاب ویژگی در مدل انجام می‌دهد.

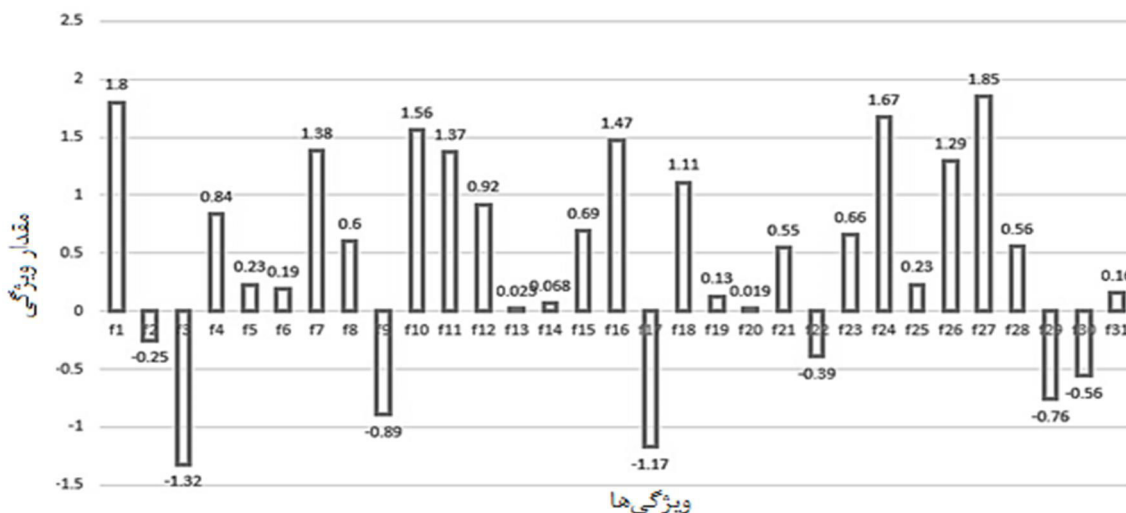
## ۴-۳-۲-۴ رگرسیون توانی و سری توانی

دو نوع دیگر از رگرسیون که جهت مدل‌سازی مورد استفاده قرار گرفت، مدل‌های توانی و سری توانی بود. در این حالت دوباره مجموعه ویژگی‌ها به دو قسمت آموزش و تست تقسیم گردید و با توجه به آزمایش‌های متعدد انجام شده، مدلی که برای داده‌ها استخراج گردید، معادلات آن‌ها برای ارزیابی نظرات به صورت زیر به دست آمد.

$$Y = X_1^{1.80} X_2^{-0.25} X_3^{-1.32} \dots X_{11}^{1.37} \dots X_{22}^{-0.39} \dots X_{31}^{0.16} \quad (1-4)$$

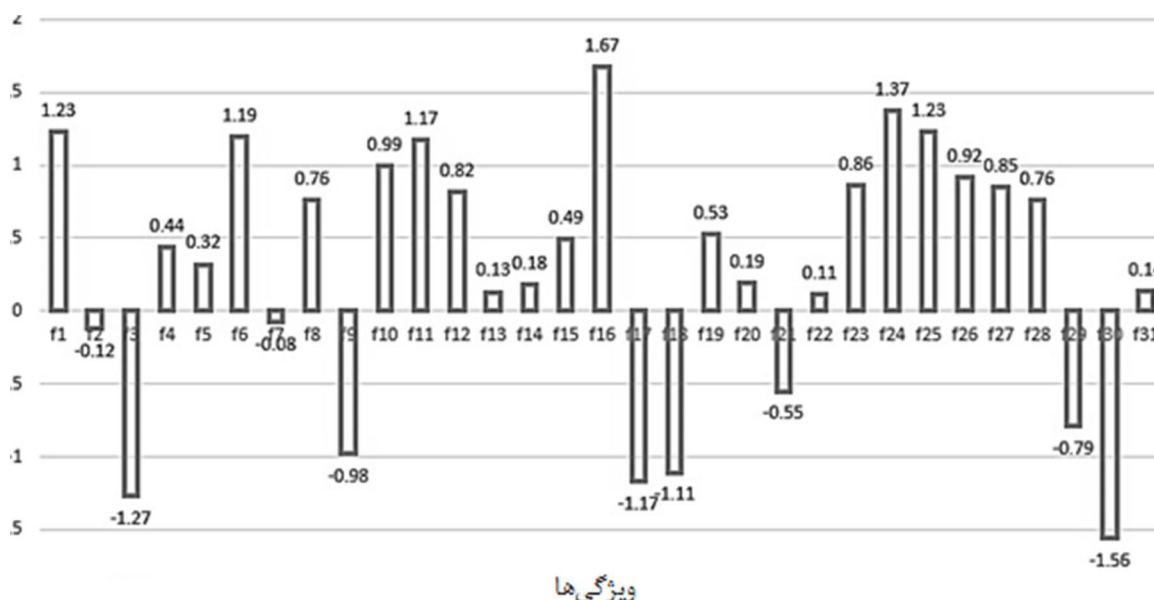
$$Y = 1.6X_1^{1.23} + 0.9X_2^{-0.12} + 0.85X_3^{-1.27} + \dots + 1.01X_{22}^{0.11} + \dots + 1.5X_{31}^{0.19} \quad (2-4)$$

با توجه به محاسبات انجام شده، ضرایب بدست آمده برای معادلات مربوط به این مدل‌ها، به ترتیب در شکل‌های ۴-۹ و ۴-۱۰ نشان داده شده است. با توجه به مقدار هر ضریب می‌توان تاثیر هر متغیر وابسته را بر متغیر مستقل بدست آورد. ارزیابی صورت گرفته بر اساس سنج‌های آماری در نظر گرفته به ترتیب برای معادلات به دست آمده،  $R^2$  برابر ۰/۴۵، RMSE برابر ۰/۲۹ و MAPE برابر ۰/۷ و برای معادله خط سری توانی مقدار  $R^2$  برابر ۰/۶۷، RMSE برابر ۰/۲۶ و MAPE برابر ۰/۶۸ حاصل گردید.



شکل ۴-۹ ضرایب معادله رگرسیون غیرخطی توانی

همانطور که از نتایج پیداست مدل سری توانی نتیجه بهتری را نسبت به مدل توانی گزارش نموده است. همچنین نسبت به دیگر روش‌ها به جز لسو نتایج بهتری را گزارش نمود که این نتایج با حدس در نظر گرفته شده در پیشنهادیه همخوانی داشت.



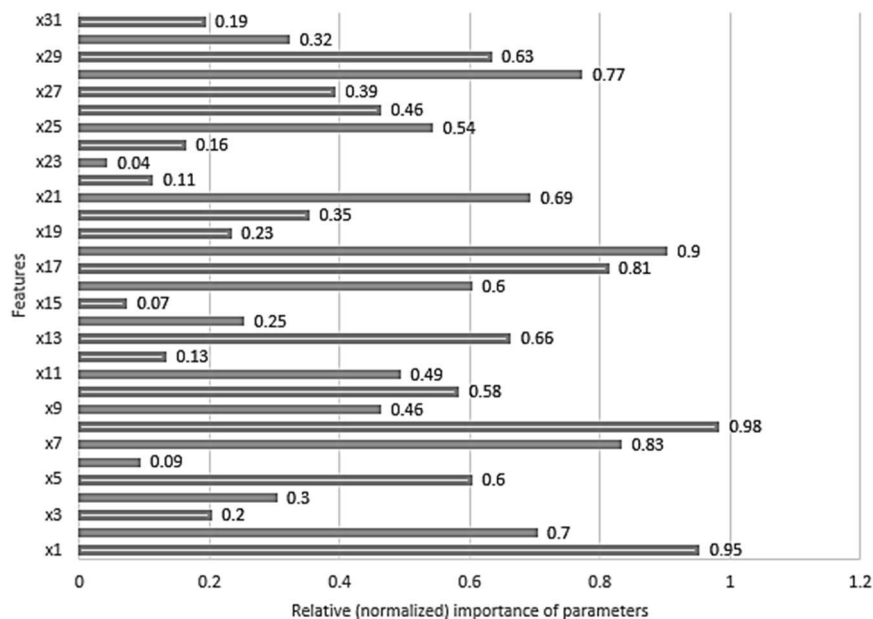
شکل ۴-۱۰ ضرایب معادله رگرسیون غیرخطی سری توانی

#### ۴-۳-۲-۵ استفاده از رگرسیون جنگل تصادفی به منظور تعیین ویژگی‌های مهم

جنگل تصادفی<sup>۱</sup> می‌تواند برای رتبه‌بندی اهمیت متغیرها در یک رگرسیون یا مشکل کلاس‌بندی به کار گرفته شود. نخستین گام در اهمیت اندازه‌گیری متغیرها در مجموعه داده، جای دادن جنگل تصادفی در داده‌ها است. بدین معنی که در هنگام انجام این فرآیند، برای جایگذاری داده‌های ناقص از میانگین‌گیری استفاده نمود. از این روش بیشتر برای بهبود دقت پیش‌بینی و کنترل استفاده می‌نمایند. اندازه زیر نمونه‌ها، همیشه همان اندازه نمونه اصلی است که با جایگزینی تنظیم می‌شوند (البته این اندازه می‌تواند توسط کاربر تغییر نماید). به طور کلی مدل‌های درخت تصمیم‌گیری و مدل‌های جنگل تصادفی برای طبقه‌بندی مورد استفاده قرار می‌گیرند. با این حال، ایده جنگل تصادفی به عنوان یک تخمین‌زننده بر روی درخت تصمیم‌گیر برای حل مشکلات رگرسیونی، در

<sup>1</sup> Random Forest

پژوهش‌های صورت گرفته توسط دیگر محققین، نتایج قابل قبولی از خود نشان داده است. در این آزمایش نیز برای انتخاب ویژگی‌های تاثیرگذار در مدل‌سازی از رگرسیون جنگل تصادفی بهره گرفته شد. نتایج حاصل از شبیه‌سازی این نوع از رگرسیون در شکل ۴-۱۱ نمایش داده شده است. با هر بار اجرای این نوع رگرسیون، نتیجه متفاوتی ایجاد خواهد شد اما در اغلب موارد، درخت تصمیم‌گیری واحدی (که به احتمال زیاد ویژگی‌های نادرست را یاد می‌گیرد) ایجاد می‌گردد. این بدین معناست که هر بار به ویژگی‌های که دارای اهمیت کمتری هستند ضریب کمتری تخصیص می‌دهد. همانطور که در شکل نشان داده شده مقدار ضرایب تاثیر بین بازه صفر تا یک گزارش شده است. برای بعضی از ویژگی‌ها این مقدار نزدیک به یک و برای برخی نزدیک به صفر می‌باشد که این بدین معناست که هر عدد بدست آمده برای هر ویژگی نشان دهنده تاثیر آن بر مدل بدست آمده است. بنابراین ویژگی با ضریب تاثیر بزرگتر به معنی اثرگذاری بیشتر می‌باشد. با در نظر گرفتن این اعداد و با اعمال آستانه ۰/۵ برای انتخاب ویژگی‌های با تاثیرگذاری بیشتر، از بین رگرسیون‌های تست شده، دو رگرسیون لسو و سری توانی که نتیجه مناسب‌تری ارائه نموده بودند را دوباره انتخاب و این بار با استفاده از این مجموعه ویژگی برای مدل‌سازی از این رگرسیون‌ها استفاده گردید.



شکل ۴-۱۱ تعیین ضریب تاثیر هر ویژگی با استفاده از رگرسیون جنگل تصادفی



نتایج بدست آمده حاکی از دقیق‌تر شدن مدل و کمتر شدن خطا بود. بر اساس معیارهای ارزیابی، نتایج بدست آمده برای رگرسیون لسو،  $R^2$  برابر ۰/۷۸، RMSE برابر ۰/۱۹ و MAPE برابر ۰/۴۷٪ و برای رگرسیون سری توانی  $R^2$  برابر ۰/۷۳، RMSE برابر ۰/۲۱ و MAPE برابر ۰/۵٪ حاصل گردید.

#### ۴-۳-۲-۶ انتخاب ویژگی

مساله انتخاب ویژگی، یکی از مسائلی است که در شناسایی آماری الگو مطرح است. این مساله در طبقه‌بندی، استخراج مدل و دیگر موارد مسائل بینایی ماشین اهمیت به سزائی دارد. در مسائل مختلف، معمولاً تعداد زیادی ویژگی وجود دارد، که بسیاری از آن‌ها یا بلااستفاده هستند و یا اینکه بار اطلاعاتی چندانی ندارند. حذف نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد مورد نظر بالا می‌برد. در پروسه انتخاب ویژگی، الگوریتم‌های متعددی وجود دارد که متناسب با کار مورد نظر می‌توان از آن‌ها استفاده نمود. الگوریتم‌های انتخاب ویژگی به دو دسته فیلتر و رپر<sup>۱</sup> تقسیم می‌شوند. روش‌های مربوط به دسته فیلتر، مستقل از طبقه‌بند هستند و از نظر محاسباتی کارا می‌باشند. اما در روش‌های مبتنی بر رپر از خروجی طبقه‌بندی کننده جهت انتخاب ویژگی استفاده می‌کنند. اگرچه روش رپر نسبت به روش فیلتر هزینه‌های محاسباتی بالاتری دارد اما از دقت بالاتری برخوردار هستند. از بین روش‌های موجود برای انتخاب ویژگی، روش انتخاب ویژگی به صورت بازگشتی حذف شونده ( $RFE^2$ ) انتخاب گردید. دلیل عمده آن را می‌توان مقیاس‌پذیری این روش و سادگی استفاده و محاسبات سریع آن دانست. در روش  $RFE$  تعداد ویژگی‌ها بصورت بازگشتی کاهش یافته و در هر مرحله دقت طبقه‌بند (در اینجا SVM یا KNN3) محاسبه و ویژگی‌هایی که بالاترین دقت را در طبقه‌بندی ایجاد می‌کنند، انتخاب می‌شوند. ملاک انتخاب ویژگی‌ها وزن آن‌ها است. بدین صورت که ابتدا برای کل ویژگی‌های موجود، طبقه‌بند آموزش داده و بردار وزن محاسبه می‌گردد. سپس بر این اساس، ویژگی‌ها مرتب شده و دقت طبقه‌بند محاسبه می‌شود. یک

<sup>1</sup> Wrapper

<sup>2</sup> Recursive feature elimination

ویژگی با پایین‌ترین وزن از مجموعه ویژگی‌ها حذف و دقت طبقه‌بند با بقیه ویژگی‌ها محاسبه می‌گردد. این روند ادامه می‌یابد تا تنها یک ویژگی باقی بماند. در نهایت با بررسی میزان درستی طبقه‌بندی که در هر مرحله محاسبه شد، می‌توان نتیجه‌گیری کرد که با چه میزان ویژگی، بیشترین دقت حاصل می‌شود. بعد از اعمال این روش، تعداد ۱۹ ویژگی باقی ماندند که ویژگی‌های  $X_1, X_2, X_3, X_7, X_8, X_{10}, X_{11}, X_{12}, X_{15}, X_{17}, X_{18}, X_{21}, X_{23}, X_{25}, X_{26}, X_{27}, X_{28}, X_{29}, X_{31}$  می‌باشند.

در ادامه در تلاشی دوباره، با استفاده از رگرسیون‌های لسو و سری توانی به مدل‌سازی با استفاده از این ویژگی‌ها پرداختیم که بهترین مدل کسب شده بر اساس معیارهای ارزیابی، برای رگرسیون لسو،  $R^2$  برابر ۰/۸۱، RMSE برابر ۰/۱۶۵ و MAPE برابر ۳/۹٪ و برای رگرسیون سری توانی  $R^2$  برابر ۰/۷۹، RMSE برابر ۰/۱۸ و MAPE برابر ۴/۵٪ حاصل گردید.

#### ۴-۴- آزمایش شماره ۳: ارزیابی ویژگی‌های پیشنهادی

همانطور که در آزمایش دوم توضیح داده شد. برای یافتن تاثیر هر ویژگی از رگرسیون جنگل تصادفی استفاده گردید. از آنجایی که در پایگاه داده تهیه شده، دو مجموعه داده وجود داشت. ماتریس ویژگی به صورت جداگانه برای هر دو مجموعه محاسبه و سپس میزان تاثیر هر یک از ویژگی‌ها در مجموعه‌هایشان محاسبه گردید. نتایج حاصل، نشان دهنده تفاوت بین ویژگی‌های تاثیرگذار بود که دلیل این امر بهینه بودن مجموعه دوم پایگاه داده نسبت به مجموعه اول بود. بنابراین، به منظور اندازه‌گیری میزان تأثیر ویژگی‌های پیشنهادی در ارزیابی سیستم‌های IQA، همبستگی بین نمرات داده شده توسط انسان به مکالمه‌ها و نتایج حاصل محاسبه گردید. به عبارتی میزان همبستگی بین هر ویژگی و خروجی مدل محاسبه گردید. همان‌طور که قبلاً ذکر شد، برای محاسبه ضریب همبستگی از دو پایگاه داده که یکی حاصل از تعامل با چهار سیستم IQA و دیگری حاصل از مکالمات روزمره بود استفاده نمودیم.

جدول ۴-۱۰ ضرایب همبستگی پیرسون برای پایگاه داده ایجاد شده

ویژگی‌ها	پایگاه داده خروجی IQA ها	پایگاه داده مکالمات
ویژگی ۱	۰/۹۲	۰/۹۱
ویژگی ۲	۰/۹۴	۰/۹۱
ویژگی ۳	۰/۸۵	۰/۸۴
ویژگی ۴	۰/۷۵	۰/۷۵
ویژگی ۵	۰/۷۸	۰/۷۳
ویژگی ۶	۰/۶۸	۰/۵۸
ویژگی ۷	۰/۹۱	۰/۹۰
ویژگی ۸	۰/۸۹	۰/۹۱
ویژگی ۹	۰/۷۳	۰/۷۸
ویژگی ۱۰	۰/۵۸	۰/۵۹
ویژگی ۱۱	۰/۶۳	۰/۷۲
ویژگی ۱۲	۰/۷۲	۰/۷۹
ویژگی ۱۳	۰/۵۲	۰/۵۴
ویژگی ۱۴	۰/۶۱	۰/۶۴
ویژگی ۱۵	۰/۸۴	۰/۸۴
ویژگی ۱۶	۰/۹۴	۰/۷۹
ویژگی ۱۷	۰/۹۰	۰/۸۶
ویژگی ۱۸	۰/۸۲	۰/۹۱
ویژگی ۱۹	۰/۶۹	۰/۷۸
ویژگی ۲۰	۰/۵۵	۰/۶۳
ویژگی ۲۱	۰/۷۸	۰/۸۱
ویژگی ۲۲	۰/۷۴	۰/۶۶
ویژگی ۲۳	۰/۶۹	۰/۷۱
ویژگی ۲۴	۰/۸۵	۰/۷۹
ویژگی ۲۵	۰/۷۰	۰/۸۱
ویژگی ۲۶	۰/۹۴	۰/۷۸
ویژگی ۲۷	۰/۹۳	۰/۶۸
ویژگی ۲۸	۰/۸۴	۰/۵۶
ویژگی ۲۹	۰/۹۲	۰/۸۸
ویژگی ۳۰	۰/۸۹	۰/۹۱
ویژگی ۳۱	۰/۸۱	۰/۷۶

نتایج حاصل از محاسبه ضریب همبستگی پیرسون برای مقدار پیش‌بینی شده نظرات در مقابل نظرات انسانی هر دو پایگاه داده در جدول ۴-۱۰ نشان داده شده است. همانطور که در جدول نشان داده شده، ۳۱ ویژگی حاصل از شانزده رابطه تعریف شده در مرحله استخراج ویژگی برای محاسبه همبستگی آن با خروجی در نشر گرفته شده است و میزان همبستگی بین هر ویژگی با دیگر ویژگی‌ها در اینجا گزارش نشده است. اما با بررسی که بر روی میزان همبستگی ویژگی‌ها صورت گرفت مقدار بدست آمده برای ضرایب همبستگی نزدیک به صفر بود. بنابراین می‌توان روند انتخاب ویژگی را برای این مجموعه داشته باشیم. با توجه به این جدول می‌توان گفت که، همه ویژگی‌ها همبستگی مثبت و قوی‌ای با نظرات انسانی موجود در پایگاه داده دارند (بهترین مقادیر حاصل در هر ستون با رنگ متفاوتی مشخص شده است). همان‌طور که از نتایج پیدا است. ویژگی‌های با شماره ۱، ۲، ۳، ۷، ۸، ۱۵، ۱۷، ۱۸، ۲۱، ۲۵، ۲۹ و ۳۱ دارای مقدار بیشتر از ۰/۸ در مجموعه پایگاه داده مکالمات می‌باشند که نشان از تأثیر بیشتر این ویژگی‌ها در مدل‌سازی برای نظرات انسانی است. برای بررسی این موضوع، مکالمات را در دو گروه با امتیاز خوب و بد در نظر گرفته و با استفاده از این ویژگی‌ها به دسته‌بندی مکالمات پرداختیم. از بین دسته‌بندهای مختلف، دسته‌بند  $SVM^1$  بهترین پاسخ رانشان داد. بر اساس معیار F عدد بدست آمده برابر با مقدار ۰/۸۹ حاصل گردید.

#### ۴-۵- آزمایش شماره ۴: استخراج مدل رگرسیونی به کمک برنامه‌نویسی بیان

##### ژن

به دلیل تعداد ویژگی بالا، جهت به دست آوردن بهترین مدل نیاز به سعی و خطا در مورد نوع معادله است که این یکی از ضعف‌های کار انجام شده محسوب می‌گردد. بنابراین با استفاده از روش‌های کاهش ویژگی و در ادامه به کمک روش برنامه‌نویسی ژن به رفع این ضعف پرداختیم. در این آزمایش تعداد ۳۶۰ مکالمه از پایگاه داده را به صورت تصادفی انتخاب و ویژگی‌ها انتخاب شده از آن‌ها

<sup>1</sup> Support vector machine

استخراج گردید و بر اساس آن ماتریس ویژگی شکل گرفت. سپس با در نظر گرفتن مقادیر اولیه برای رسیدن به بهترین مدل از GEP استفاده نمودیم. مقادیر اولیه برای استفاده از بیان ژن در جدول ۴-۱۱ نمایش داده شده است. برای به دست آوردن بهترین مدل توسط برنامه‌نویسی بیان ژن از نرم‌افزار GeneXpro نسخه ۵ استفاده نمودیم. معیار ارزیابی برای یافتن بهترین مدل را RMSE در نظر گرفتیم. در برنامه‌نویسی بیان ژن معمولاً شرط توقف (خاتمه) در دو حالت می‌توان تعیین نمود. یکی معیار ارزیابی در نظر گرفته شده که تا رسیدن به کمترین مقدار برنامه پایان یابد و دیگری شرط تعداد تکرار در صورت نرسیدن به مقدار مناسب معیار، که در شرایط شبیه‌سازی انجام شده مقدار ماکزیمم تکرار را ۱۰۰۰ فرض کردیم تا در صورتی که شرط اول محقق نگردد توسط این شرط برنامه خاتمه یابد. شرایط دیگری که در طول اجرا در نظر گرفتیم در جدول ۴-۱۲ آورده شده است. مجموعه آموزش به میزان ۷۰ درصد و مجموعه آزمون به میزان ۳۰ درصد در نظر گرفته شد. با توجه به شرایط در نظر گرفته شده مدل استخراج گردید. معادله خط به دست آمده در فرمول ۴-۲ نمایش داده شده است. همانطور که اشاره شد معادله خط در نظر گرفته شده معادله‌ای است که بر روی ۳۶۰ نمونه در نظر گرفته شده منطبق گردیده است. بنابراین با داشتن این معادله می‌توانیم برای یک مکالمه دیده نشده ابتدا استخراج ویژگی انجام داده، سپس با قرار دادن مقادیر بدست آمده برای هر ویژگی در معادله، عدد حاصل از خروجی این معادله را بدست آورد. انتظار می‌رود که عدد بدست آمده کمترین اختلاف نسبت به عددی که مد نظر کاربر بوده را داشته باشد. مقدار RMSE در این حالت برابر ۰/۱۰۵ حاصل گردید. همچنین ضریب  $R^2$  برابر با ۰/۸۵ حاصل گردید که بیانگر قابل قبول بودن نتیجه (معادله خط رگرسیونی) به دست آمده است. در ادامه با بررسی روی داده‌ها و با توجه به امتیازات داده شده توسط کاربران به هر مکالمه این فرض در نظر گرفته شد که می‌توان مکالمات را به سه دسته با امتیازات خوب، متوسط و بد در نظر گرفت سپس برای هر دسته از این مکالمات یک معادله خط به دست آورد تا بتوان حدس آگاهانه‌تری نسبت به امتیاز داده شده توسط کاربر داشت. لذا از برای پیاده‌سازی این فرض از روش K-means برای دسته‌بندی مکالمات استفاده نمودیم. که دقت

۹۴/۳ درصدی را برای دسته‌بندی مکالمات به همراه داشت و توانست به خوبی مکالمات را در این سه دسته قرار دهد. بنابراین این فرآیند می‌توانست معادله خط دقیق‌تری برای هر کلاس بدست آورد. پس با در نظر گرفتن هر دسته از مکالمات و ساختن مجموعه آموزش و آزمون به ترتیب به میزان ۷۰ و ۳۰ درصد معادله خط برای هر کلاس استخراج گردید.

جدول ۴-۱۱ مقادیر اولیه برای برنامه‌نویسی بیان ژن

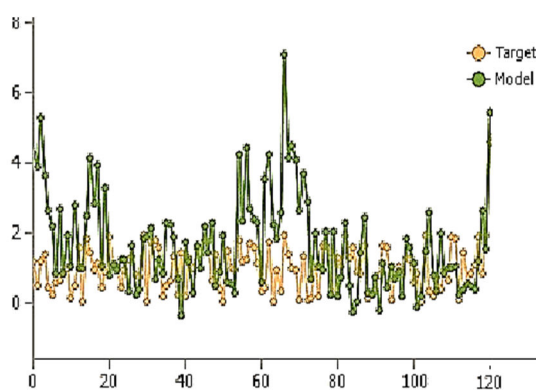
مقدار	متغیر	ردیف	مقدار	متغیر	ردیف
۰/۰۰۱۳۸	Mutation	۶	۰/۰۰۲۷۷	Gene transposition	۱
۰/۰۰۵۴۶	Permutation	۷	۰/۰۰۲۶	Random chromosomes	۲
۰/۰۰۵۴۶	Biased mutation	۸	۰/۰۰۱۰۲	Random cloning	۳
۰/۰۰۵۴۶	Inversion	۹	۰/۰۰۲۶	Best cloning	۴
۰/۰۰۵۴۶	Tail mutation	۱۰	۰/۰۰۲۷۷	Gene recombination	۵

جدول ۴-۱۲ شرایط شبیه‌سازی به کمک برنامه‌نویسی بیان ژن

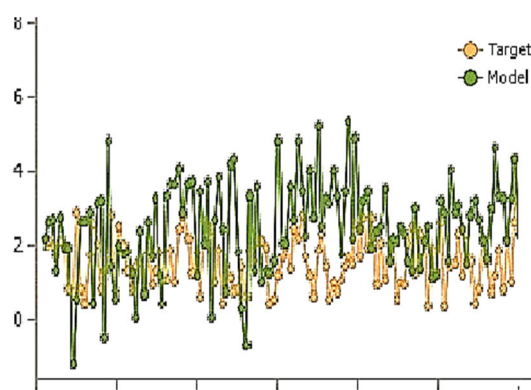
مقدار	پارامتر	ردیف
۱۰	Number of function	۱
۲۵۲	Train record	۲
۱۰۸	Test record	۳
۳۰	Number of chromosomes	۴
۱۰	Head size	۵
۴	Number of Genes	۶
Addition	Linking Function	۷
RMSE	Fitness function	۸
Max fitness	Stop condition	۹

$K1 = 1.659797;$   
 $K2 = 4.165807;$   
 $K3 = -7.002964;$   
 $K4 = 4.934577;$   
 $K5 = -1.55237;$   
 $K6 = 6.148657;$   
 $K7 = -3.034376;$   
 $K8 = 5.689516;$   
 $Y = (\exp(\sin(d(20), d(19), K8), \min(d(10), K4, d(15)),$   
 $\max(d(1), K4, K6)) * \exp(\text{gep3Rt}(d(12), d(18), K7),$   
 $\min(d(11), K7, \tanh(d(20)), \sin(d(16), K5, d(2), d(1)))) +$   
 $(\min(d(11), (\text{gep3Rt}((1.0 - \text{reallog}((1.0 / (d(14)))))) +$   
 $d(13))) + d(14)) + \max(((d(20) - d(4)) + \text{reallog}(d(20)))$   
 $+ K5), (((d(16) - d(10)) - d(15)) + \text{gep3Rt}(d(17)))) +$   
 $(\exp((d(8) + \text{gep3Rt}(((1.0 / (d(15))) * d(14)) - (d(10)^2)))) + d(14));$   
 $\text{function result} = \text{gep3Rt}(x)$   
 $\text{if } (x < 0.0),$   
 $\text{result} = -((-x)^{1.0 / 3.0});$   
 $\text{else}$   
 $\text{result} = x^{(1.0 / 3.0)};$   
 $\text{end}$

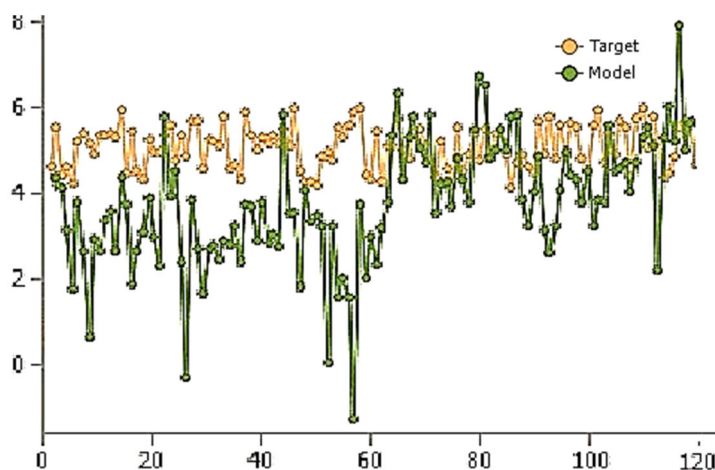
شکل ۴-۱۲ خط رگرسیون به دست آمده برای هر کلاس و معادله ۳-۴، رابطه این معادلات خط برای هر کلاس را نمایش می‌دهد. مقدار به دست آمده با توجه به معیار ارزیابی RMSE برابر ۰/۰۹ حاصل گردید.



(ب)



(الف)



(ج)

شکل ۴-۱۲ الف) معادله خط به دست آمده برای مکالمات مربوط به کلاس خوب ب) معادله خط به دست آمده برای مکالمات مربوط به کلاس متوسط ج) معادله خط به دست آمده برای مکالمات مربوط به کلاس ضعیف

همانطور که در شکل مشاهده می‌نمایید مدل بدست آمده به خوبی هدف را دنبال می‌نماید. نتایج حاصل حاکی از دقت مناسب مدل می‌باشد. برای ارزیابی معادلات بدست آمده مقدار  $R^2$  برای هر یک از معادلات خط به دست آمد که این مقادیر به ترتیب برابر  $0/91$ ،  $0/89$  و  $0/94$  حاصل گردید که بیانگر قابل قبول بودن نتایج به دست آمده برای هر کلاس است. معادلات خط رگرسیونی بدست آمده برای هر کلاس به ترتیب برای مکالمات کلاس با امتیاز خوب، متوسط و ضعیف در معادلات ۳-۴ تا ۵-۴ آورده شده است.

$$K1 = 1.36878780480361;$$

$$K2 = 9.37502670369579;$$

$$K3 = -10.0029402895086;$$

$$y = ((K1 + d(14)) - (\min(d(10), \tanh(((d(17) + d(20)) / 2.0))) + d(10))) + (\min(d(11), (\text{gep3Rt}((1.0 - \text{reallog}((1.0 / (d(14))))))) + d(13))) + d(14)) + (((((d(16) + d(16)) - (d(15) * d(13))) + ((K2 - 3) * d(16))) + (d(20) * d(13))) / 2.0) + (\text{exp}((d(8) + \text{gep3Rt}(((1.0 / (d(15)))) * d(14)) - (d(10)^2)))))) + d(14));$$

۳-۴



$$\begin{aligned}
K1 &= 0.579841168210461; \\
K2 &= 4.86109679354414; \\
K3 &= -0.824718926398679; \\
K4 &= -3.91473723716566; \\
K5 &= -8.59184545426801; \\
K6 &= 4.44166217688528; \\
K7 &= 11.3615450430273; \\
K8 &= 0.74112286568496; \\
K9 &= -10.6840849162225; \\
K10 &= 9.62279122287667;
\end{aligned}$$

(۴-۴)

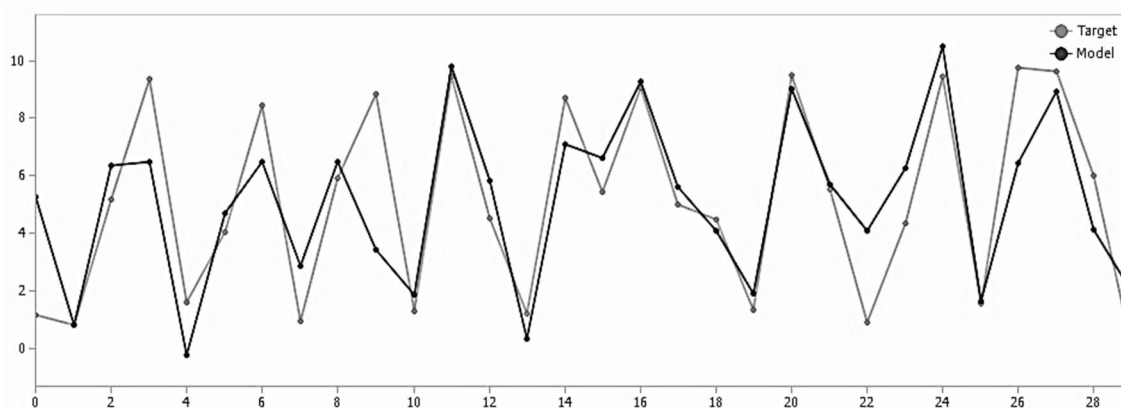
$$\begin{aligned}
y &= \left( \left( d(20) * (\min(K1, d(13)) * ((K2 + G1C5) / 2.0)) \right) + \right. \\
&\quad \left. \left( ((1.0 - K4) + d(14)) + d(14) \right) \right) \\
&+ \max \left( \left( ((d(20) - d(4)) + \text{reallog}(d(20))) + K5 \right), \right. \\
&\quad \left. \left( ((d(16) - d(10)) - d(15)) + \text{gep3Rt}(d(17)) \right) \right) + \\
&+ \left( \left( \tanh((d(16) * K6)) * (\text{atan}(d(14)) * (K7 + d(15)))) \right) + \right. \\
&\quad \left. \left( (d(12) * d(12)) - (d(19) + d(19)) \right) \right) \\
&\left( d(14) - \left( K8 * \left( K9 + \left( \left( (d(15) + d(10)) / 2.0 \right) + d(8) \right) * (K10 - K9) \right) \right) \right);
\end{aligned}$$

$$\begin{aligned}
K1 &= 2.42611828071061; & K2 &= -0.705328684210632; \\
K3 &= -2.69619839522557; & K4 &= -0.566471205090578; \\
K5 &= -0.552774197378301; & K6 &= 9.15565779946092; \\
K7 &= 4.06337613705845; & K8 &= 5.78551679530081; \\
K9 &= -3.82179032746965; & K10 &= -0.360356306526117; \\
K11 &= 1.05161744968877; & K12 &= 1.77236012189421; \\
K13 &= 8.93775885461439; & K14 &= -8.94146549882504; \\
K15 &= 5.34362234380226e-02; & K16 &= 2.39474081322003; \\
K17 &= -6.24432169771755; & K18 &= -2.8893895483539;
\end{aligned}$$

$$Y = K8 + \left( \begin{array}{l} \exp \left( \begin{array}{l} \sin(K9, d(19), d(13), K14), d(K10, d(20), d(19), d(19)), \\ \min(d(3), d(11), K2, d(20)) \end{array} \right), \\ \tanh(\min(d(14), d(19)), \max(d(13), K14, K2), d(14), d(20)), \\ \tanh(K1, K2, K4, d(20)), \sin(d(20), d(14), d(14), d(13)) \end{array} \right) + \\ \exp \left( \begin{array}{l} ((\min(d(10), K5, d(4)) + \text{gepAMin2}(d(19), d(16))) / 2.0), \\ \max \left( \begin{array}{l} d(d(19), K11, d(15)), \\ \text{gep3Rt}(d(17)), \tanh(d(15), d(10)) \end{array} \right), \\ \text{gepAMin2}(\min(d(17), d(18), K5, d(6)), d(2)) \end{array} \right) + \quad (5-4)$$

$$\left( \begin{array}{l} \exp \left( \begin{array}{l} \sin(d(20), d(19), K12), \min(d(10), K13, d(15)), \\ \max(d(1), K13, K6) \end{array} \right) * \\ \exp \left( \begin{array}{l} \text{gep3Rt}(d(12), d(18), K7), \min(d(11), K14), \\ \tanh(d(20)), \sin(d(16), K15, d(2), d(1)) \end{array} \right) \end{array} \right) \\ + \exp \left( \begin{array}{l} \sin(\min(d(13), d(13), K16)), \\ \tanh \left( \begin{array}{l} \min(K17, d(19), K18, d(15)), \\ \text{gep3Rt}(d(17), d(13), d(14), d(20)), \sin(d(20), K8) \end{array} \right), \\ \min(\exp(d(13)), \max(d(12), d(6)), d(20)) \end{array} \right);$$

جهت بررسی بهتر معادلات بدست آمده، مکالمات جدیدی در نظر گرفته شد و بعد از تخصیص دسته‌بندی و تخصیص به یک کلاس، امتیاز مکالمه محاسبه گردید. نتایج حاصل از دقت بالای معادلات بدست آمده بود. شکل ۴-۱۳ یک نمونه از تصدیق مجموعه آزمون از داده‌ها را بر روی مدل به دست آمده برای یک کلاس نمایش می‌دهد.



شکل ۴-۱۳ تصدیق مدل به دست آمده برای کلاس خوب با استفاده از داده‌های آزمون

بنابراین با داشتن سه معادله خط بالا برای کلاس‌های خوب، متوسط و ضعیف می‌توان به راحتی برای هر مکالمه جدید صورت گرفته با سیستم و بر اساس ویژگی‌های استخراج شده از مکالمه، امتیاز مدنظر کاربر را پیش‌بینی نمود.

جدول ۴-۱۳ مقادیر به دست آمده برای پیش‌بینی نظرات کاربران به تعداد ۵ نمونه برای هر کلاس

X1	X2	X3	X4	X7	X9	X10	X12	X15	X16	X17	X19	Human score	Predicted score
0.060 606	0.030 909	0.037 411	0.088 889	0.047 13	0.055 333	0.287 249	0.517 241	0.081 481	0.000 1	0.000 1	6.406 429	1.9	1.9
0.021 739	0.031 954	0.038 205	0.033 333	0.046 72	0.053 24	0.191 201	0.481 481	0.011 111	0.000 1	0.000 1	6.500 11	1.2	0.99
0.048 193	0.029 094	0.034 875	0.076 923	0.047 542	0.055 57	0.314 884	0.461 538	0.100 962	0.000 1	0.000 1	5.468 808	1.3	1.27
0.103 896	0.030 647	0.033 406	0.166 667	0.055 785	0.060 279	0.319 372	0.383 333	0.098 958	0.000 1	0.000 1	5.549 738	1.25	1.12
0.148 936	0.068 408	0.049 792	0.259 259	0.138 663	0.107 248	0.684 211	0.507 353	0.101 852	0.021 739	0.013 158	5.675 748	0.58	0.55
0.086 614	0.035 063	0.031 453	0.132 53	0.058 389	0.052 293	0.364 2	0.406 061	0.120 482	0.015 385	0.005 479	2.046 81	5.2	5.00
0.062 992	0.042 358	0.039 908	0.086 022	0.059 018	0.054 769	0.367 355	0.524 0524	0.109 677	0.004 545	0.000 1	2.104 353	4.5	5.00
0.072 993	0.041 097	0.038 004	0.102 041	0.060 026	0.054 986	0.469 125	0.474 576	0.114 286	0.006 452	0.000 1	2.168 761	5.6	5.5
0.164 062	0.039 763	0.037 068	0.212 121	0.062 695	0.058 162	0.302 264	0.374 648	0.159 596	0.004 255	0.000 1	2.284 189	5.9	6.00
0.092 308	0.048 631	0.046 782	0.125	0.070 133	0.066 656	0.475 309	0.537 931	0.15	0.010 989	0.004 651	2.359 473	5.4	5.5
0.105 769	0.059 911	0.055 173	0.152 778	0.092 628	0.084 953	1.024 216	0.922 222	0.119 444	0.023 881	0.012 903	5.913 188	9.7	9.5
0.098 361	0.053 15	0.049 404	0.15	0.087 118	0.080 93	0.737 586	0.721 212	0.1	0.026 667	0.011 429	6.014 084	8.9	8.5
0.074 468	0.042 835	0.035 557	0.112 903	0.067 835	0.056 178	0.619 003	0.717 391	0.077 419	0.007 018	0.000 1	6.109 227	8.7	8.8
0.163 265	0.040 563	0.034 625	0.213 333	0.064 319	0.054 801	0.527 019	0.694 545	0.152	0.008 571	0.000 1	6.242 091	9.4	9.10
0.099 237	0.047 696	0.035 725	0.141 304	0.073 572	0.055 706	0.478 459	0.563 077	0.102 174	0.006 897	0.000 1	6.349 994	9.4	9.5

همچنین برای درک بهتر نسبت به کارایی مدل پیشنهادی در پیش‌بینی نظرات، تعدادی از امتیازات داده شده توسط کاربران و امتیاز به دست آمده از مدل به همراه مقادیر برخی از ویژگی‌ها در جدول ۴-۱۳ نمایش داده شده است. بنابراین نتایج به دست آمده از آزمایش‌های مختلف نشان داد که مدل به دست آمده از کارایی مناسبی برخوردار است. همچنین روش پیشنهادی در کارهای مشابه قبلی مشاهده نگردید که این یکی دیگر از نوآوری‌ها صورت گرفته در این پژوهش است. در محاسبه امتیاز یک مکالمه جدید، باید توجه نمایید که در ابتدا استخراج ویژگی باید صورت پذیرد. سپس بر اساس میزان تعلق به هر دسته از مکالمات محاسبه و تخصیص آن به یک کلاس، از معادله خط مربوط به آن استفاده تا بتوان امتیاز داده شده توسط کاربر را پیش‌بینی نمود.

## ۴-۶ آزمایش شماره ۵: استفاده از شبکه LSTM در یادگیری عمیق جهت

### پیش‌بینی امتیاز کاربر

یکی از روش‌هایی که امروزه در تحقیقات مورد توجه قرار گرفته است استفاده از یادگیری عمیق در کاربردهای مختلف هوش مصنوعی است. شبکه‌های LSTM<sup>۱</sup>، نوع خاصی از شبکه‌های عصبی بازگشتی با توانایی یادگیری وابستگی‌های بلند مدت می‌باشند. طراحی ساختار این شبکه به صورتی است که اطلاعات خیلی دور را به خوبی یاد می‌گیرند که این ویژگی در ساختار آن‌ها نهفته است. تمامی شبکه‌های عصبی بازگشتی به شکل زنجیره‌ای تکرار شونده از واحدهای شبکه‌های عصبی هستند. در شبکه‌های عصبی بازگشتی استاندارد، این واحدهای تکرار شونده ساختار ساده‌ای دارند، برای مثال تنها شامل یک لایه tanh هستند. در حالیکه شبکه‌های LSTM نیز چنین ساختار دنباله یا زنجیره ماندنی دارند ولی واحد تکرار شونده، ساختار متفاوتی دارد. یعنی به جای داشتن تنها یک لایه شبکه عصبی، ۴ لایه دارند که طبق ساختار ویژه‌ای با یکدیگر در تعامل و ارتباط هستند. بنابراین ما از این شبکه برای مدل‌سازی رگرسیونی با استفاده از روش پنجره استفاده نمودیم. یعنی مسئله رگرسیون را

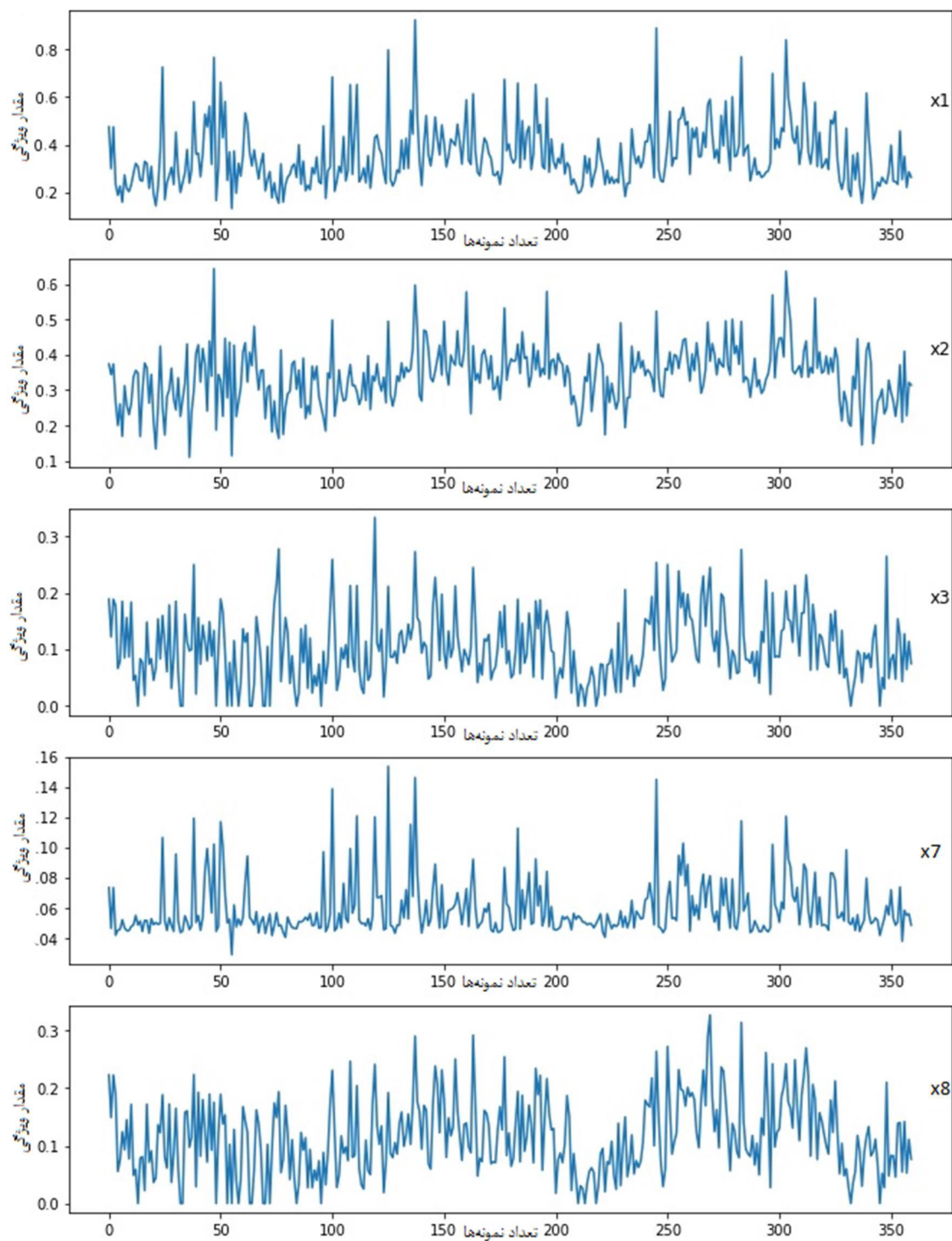
<sup>۱</sup> Long Short Term Memory

به این صورت تعریف نمودیم که با داشتن مقادیر ویژگی‌ها بتوانیم امتیاز نهایی را پیش‌بینی نماییم. به طور مثال در یک مسئله سری زمانی ما با داشتن مقدار در زمان  $t$ ،  $t-1$  و  $t-2$  می‌توانیم مقدار  $t+1$  را پیش‌بینی نماییم. لذا از این ایده برای حدس مقدار امتیاز داده شده توسط کاربران استفاده گردید. همانطور که قبلاً اشاره گردید یکی از معضلات روش‌های ارزیابی که از ارزیاب انسانی در سیستم‌های پرسش و پاسخ بهره می‌گرفتند این بود که ارزش یک سیستم از هر کاربر به کاربر دیگر تغییر می‌یابد یعنی اینکه یک کاربر ممکن است بعد از استفاده از یک سیستم امتیاز بالایی را نسبت به عملکرد آن سیستم تخصیص دهد و کاربر دیگر برای همان سیستم امتیاز پایینی را در نظر بگیرد. بدین منظور سعی گردید که با استفاده از یادگیری عمیق به پیش‌بینی امتیاز در نظر گرفته شده توسط هر کاربر برای یک سیستم پردازیم. همانطور که اشاره شد اگر ما مقدار امتیاز را در زمان  $t$ ،  $t-1$  و  $t-2$  داشته باشیم می‌توانیم مقدار امتیاز را در  $t+1$  پیش‌بینی نماییم. بنابراین فرض نمودیم که یک مکالمه در زمان  $t$  با سیستم اتفاق افتاده و مکالمات دیگر با همان سیستم در زمان‌های  $t-1$ ،  $t-2$  تا  $t-n$  رخ داده است به عبارتی سعی گردید که مسئله به شکل یک سری زمانی در نظر گرفته شود. با توجه به اینکه تعداد مکالمات صورت پذیرفته برای یک سیستم پرسش و پاسخ تعاملی در مجموعه داده مقدار زیادی نبود لذا برای آزمایش این روش، مجموعه داده دوم تهیه شده (مجموعه داده مکالمات انگلیسی) در نظر گرفته شد. از مجموع مکالمات تعداد ۳۶۰ مکالمه مربوط به یک موضوع خاص که هم دارای امتیاز خوب، متوسط و بد بودند، در نظر گرفته شد. سپس (همانطور که در قسمت انتخاب ویژگی اشاره شد، از مجموع ۳۱ ویژگی در نظر گرفته شده برای هر مکالمه) تعداد ۱۹ ویژگی که دارای اهمیت بیشتری نسبت به دیگر ویژگی‌ها بودند انتخاب و بر اساس آن‌ها از هر مکالمه استخراج و ماتریس ویژگی تشکیل گردید. در ادامه به کمک شبکه LSTM به پیش‌بینی امتیاز داده شده برای مکالمه در زمان  $t+1$  پرداخته شد. شکل ۴-۱۴ نمودار مربوط به برخی از ویژگی‌های استخراج شده برای مکالمات نشان داده شده است. جهت پیاده‌سازی شبکه LSTM از کتابخانه keras در پایتون نسخه ۳ استفاده گردید. تنظیمات شبکه LSTM بدین صورت پذیرفت که تعداد ۵۰ نرون

در لایه مخفی اول و یک نرون در لایه خروجی برای پیش‌بینی نظرات کاربران در نظر گرفته شد. جهت ارزیابی شبکه از معیار  $MAE^1$  و برای بهینه‌سازی از نسخه Adam استفاده گردید. تعداد epochها در قسمت آموزش شبکه برابر ۲۵۰ و اندازه دسته<sup>۲</sup> برابر ۷۲ در نظر گرفته شد. تنظیم این مقادیر به صورت سعی و خطا جهت رسیدن به بهترین نتیجه انجام شده است. ۷۰ درصد از نمونه‌ها جهت آموزش و ۳۰ درصد آن برای آزمون مورد استفاده قرار گرفته است. در نهایت با استفاده از مجموعه آزمون مدل مورد ارزیابی قرار گرفت و بر اساس معیار ارزیابی RMSE مقدار ۰/۰۵ و معیار صحت مقدار ۰/۵۶ حاصل گردید. که براساس معیار RMSE نسبت به روش‌های قبلی نتیجه مناسب‌تری حاصل شد ولی بر اساس معیار صحت، می‌توان به این نتیجه دست یافت که تنظیمات مناسب‌تری باید بر روی شبکه دست یافته و همچنین مجموعه داده بیشتر و مناسب‌تری مورد استفاده قرار گیرد تا بتوان مدل مناسب‌تری برای پیش‌بینی امتیاز کاربران پیشنهاد نمود. معمولاً برای بررسی و ارزیابی یک شبکه یادگیری عمیق یا از معیار دقت برای این شبکه‌ها استفاده می‌شود و یا اینکه مقادیر از دست رفته در طول دوره آموزش و تست را مورد بررسی قرار می‌دهند تا مشخص شود شبکه ایجاد شده جهت مدل‌سازی از دقت و پایداری مناسبی برخوردار است.

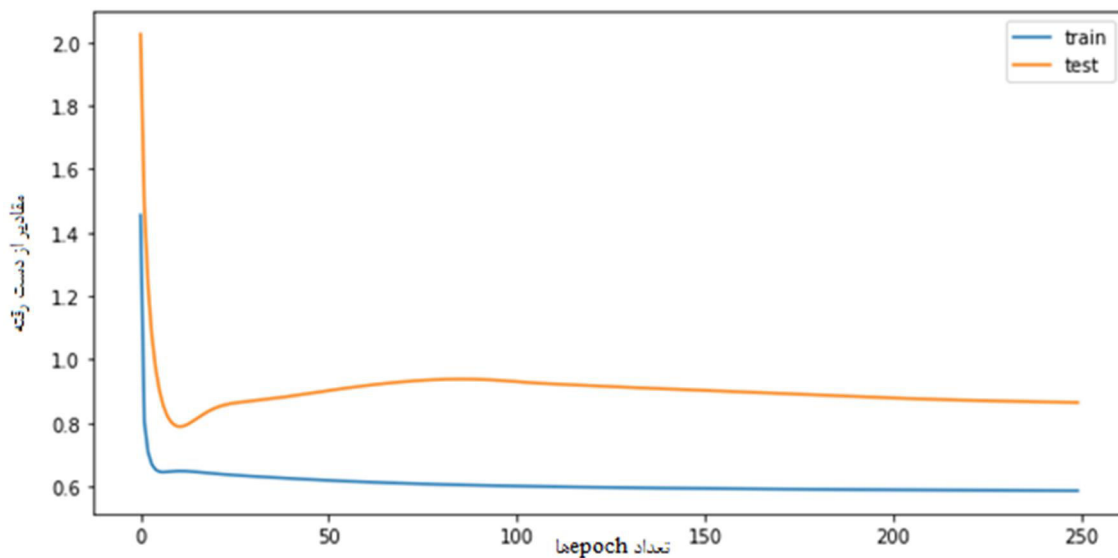
---

<sup>1</sup> Mean Absolute error<sup>2</sup> Batch Size



شکل ۴-۱۴ مقادیر برخی از ویژگی‌ها به دست آمده جهت پیش‌بینی امتیاز تخصیص شده با استفاده از LSTM

شکل ۴-۱۵ مقدار از دست رفته<sup>۱</sup> مجموعه آزمون و آموزش را برحسب تعداد epochها نمایش می‌دهد. همانطور که در این شکل نشان داده شده است بعد از epoch برابر با ۲۰ مقادیر تست و آموزش به صورت موازی با یکدیگر می‌باشند و همگرایی بین این دو شکل نمی‌گیرد. یعنی اختلاف مجموعه تست و آموزش مقدار مطلوبی نمی‌باشد. به عبارت دیگر شبکه به خوبی آموزش ندیده است تا بتوان در زمان تست آن به نتایج مطلوبی دست یافت. دلیل این امر تعداد کم مکالمات و تنظیم نادرست پارامترها می‌باشد. بدلیل اینکه برای آموزش شبکه‌های یادگیری عمیق نیاز به حجم زیادی از داده می‌باشد و در اینجا ما از این ویژگی برخوردار نبودیم به همین دلیل به نتایج مطلوبی دست پیدا نکردیم. اما با تهیه یک پایگاه داده مناسب و تنظیم پارامترها می‌توان به نتایج مطلوبی دست پیدا کرد.



شکل ۴-۱۵ نمودار مربوط به مقادیر از دست رفته مجموعه آموزش و آزمون در طول فرآیند آموزش

<sup>۱</sup> Loss



## فصل پنجم - نتیجه‌گیری و کارهای آینده

یکی از چالش‌های مورد توجه محققین از زمان ایجاد سیستم‌های پرسش و پاسخ تعاملی، ارزیابی این سیستم‌ها به منظور ارتقاء کیفیت طراحی و به روزرسانی آن‌ها بوده است. به دلیل فقدان یک روش استاندارد در زمینه ارزیابی و دخالت داشتن عوامل متعدد در ارزیابی، همواره از قضاوت‌های انسانی جهت ارزیابی این سیستم‌ها بهره گرفته شده است. در این رساله، روشی آماری جهت ارزیابی سیستم‌های پرسش و پاسخ تعاملی به منظور جایگزینی به جای ارزیاب‌های انسانی ارائه گردیده است. روشی بر اساس مکالمه صورت گرفته بین کاربر و سیستم ارائه شده تا بر اساس آن، بتوان به ارزیابی سیستم پرداخت و امتیازی که به منظور ارزیابی توسط کاربر به سیستم تخصیص داده می‌شود، را پیش‌بینی نمود. یکی از چالش‌های مهم در مدل‌سازی نظرات ارزیابی داده شده به این نوع از سیستم‌ها، این است که ارزیاب‌های انسانی از چه ویژگی‌هایی در ارزیابی استفاده و چگونه آن‌ها را با یکدیگر ترکیب می‌نمایند. در این فصل یک گزارشی مبنی بر جمع‌بندی روش پیشنهادی و جزئیات مختصری از آن اشاره شده است.

## ۵-۱ جمع‌بندی روش پیشنهادی در رساله

در این رساله یک روش آماری جهت ارزیابی سیستم‌های پرسش و پاسخ تعاملی ارائه شده است. با توجه به کارهای قبلی صورت گرفته و به دلیل وجود چالش‌های متعدد در ارزیابی سیستم‌های IQA، اکثر روش‌های موجود از روش‌های ارزیابی در سیستم‌های QA بهره گرفته‌اند که هر یک دارای اشکالات و ایرادات خاص مربوط به خود بوده و قابلیت‌های لازم به منظور استفاده به عنوان یک روش ارزیابی جامع در سیستم‌های IQA را دارا نبودند. از طرفی اتوماتیک کردن همه پارامترهای ارزیابی و ارائه یک روش همگانی امکان‌پذیر نبود. بنابراین ارزیاب‌های انسانی می‌بایست در فرآیند ارزیابی مشارکت داده شوند که این امر امکان اعمال سلیقه شخصی ارزیاب‌ها را در بر دارد. به منظور پیشگیری از این ایراد، اغلب چندین ارزیاب در طول فرآیند

ارزیابی مورد استفاده قرار می‌گرفتند تا برآیند نظرات داده شده در ارزیابی سیستم‌ها مورد توجه قرار گیرد که خود سبب صرف هزینه و زمان بسیاری در فرآیند ارزیابی‌ها می‌گردید. لذا برای استانداردسازی روش‌های ارزیابی کنفرانس‌های سالانه‌ای شکل گرفت تا براساس ساختار و چارچوب این کنفرانس‌های سالانه ارزیابی سیستم‌های IQA صورت پذیرد. اما همچنان در ارزیابی این سیستم‌ها از ارزیاب انسانی استفاده شده و تنها یک چهارچوب استاندارد جهت انجام فرآیند ارزیابی پیشنهاد شده است. روش استاندارد دی که بدون استفاده از ارزیاب انسانی به ارزیابی سیستم‌های IQA بپردازد وجود ندارد. از طرفی جایگزین کردن یک مدل کاملاً اتوماتیک که بتواند به جای انسان در کل فرآیند شرکت نماید (بدلیل عدم امکان‌پذیری در اتوماتیک کردن همه ویژگی‌های ارزیابی) مقدور نمی‌باشد ولی ارائه روشی که بتواند تاثیر انسان در روند ارزیابی را کمتر نماید، امری مهم و ضروری در راستای روال ارزیابی سیستم‌های IQA می‌باشد. لذا ارائه مدلی که بتواند در راستای تسهیل در این امر و اتوماتیک کردن فرآیند ارزیابی مورد استفاده قرار گیرد، همواره یکی از موضوعات مورد علاقه در حوزه سیستم‌های IQA می‌باشد. می‌بایستی به این نکته توجه داشت که هنوز سیستم‌های IQA استاندارد دی که به صورت ایده‌آل عمل نمایند، قابل دسترس نمی‌باشند و بیشتر سیستم‌های موجود یا صرفاً در یک زمینه خاص فعالیت نموده و یا بطور کامل (آنلاین) در اختیار کاربران نمی‌باشند. بنابراین می‌بایست یک سناریو از پیش تعیین شده در رابطه با استفاده از این سیستم‌ها برای کاربران در نظر گرفت. همچنین بدلیل عدم امکان دسترسی کامل به سیستم‌های موجود در اینترنت و صرفاً امکان مشاهده مکالمات صورت گرفته با سیستم‌ها، اساس روش پیشنهادی (که ارزیابی بر اساس آن صورت پذیرفت) بر سوال‌ها و پاسخ‌های رد و بدل شده بین کاربر و سیستم (که در

این رساله به عنوان مکالمه از آن یاد شد) صورت پذیرفت و به ناچار از در نظر گرفتن بقیه ویژگی‌های تاثیرگذار در ارزیابی این سیستم‌ها صرف نظر گردید.

یکی از مهم‌ترین مشکلات موجود جهت ارائه یک روش آماری برای ارزیابی سیستم‌های پرسش و پاسخ تعاملی عدم دسترسی کامل به این سیستم‌ها و انواع آن بود. بنابراین تنها سیستم‌های چت متن باز که زیر شاخه‌ای از این سیستم‌ها هستند، در نظر گرفته شد و بدلیل عدم وجود یک پایگاه داده استاندارد جهت ارزیابی سیستم‌های IQA، یک پایگاه داده جهت ارزیابی این سیستم‌ها ایجاد گردید. برای انجام اینکار، چند سیستم قابل دسترس، همراه با یک سیستم تعاملی پایه که در آزمایشگاه تکنولوژی زبان‌های طبیعی دانشگاه شاهرود ساخته شده بود را در نظر گرفته و با کمک دانشجویان دانشگاه و یک موسسه زبان به ایجاد این پایگاه داده پرداختیم. بر اساس مکالمات صورت گرفته کاربران با سیستم‌ها، ۶۰۰ مکالمه تولید که بر اساس نظر فرد خبره تنها ۵۴۰ مورد آن به عنوان مکالمه مناسب در نظر گرفته شد. به همراه این پایگاه داده و برای تصدیق کار انجام شده در این رساله، ۲۴۵۰ مکالمه انگلیسی با موضوعات مختلف که در آموزش زبان از آن‌ها استفاده می‌گردید نیز جمع‌آوری شد تا به عنوان یک پایگاه داده دیگر مورد استفاده قرار گیرد. در مرحله دوم از روش ارائه شده، به تعیین پارامترهایی که در فرآیند ارزیابی مورد توجه قرار می‌گیرند و امکان اتوماتیک کردن آن‌ها وجود دارد، پرداخته شد. از آنجایی که (براساس پایگاه داده شکل گرفته) تنها مکالمه‌ها و امتیاز داده شده توسط کاربران به سیستم در پایگاه داده ذخیره شده بود. ویژگی‌هایی باید مد نظر قرار می‌گرفت که بتوانند ما را به امتیاز داده شده توسط کاربران نزدیک نماید تا بر اساس آن‌ها مدلی ارائه گردد که در دفعات بعدی بتواند یک حدس آگاهانه از امتیاز داده شده توسط کاربر جدید را ارائه نماید. بنابراین به استخراج ویژگی از روی مکالمات پرداختیم. از آنجایی که در هر

کار پردازش متن در ابتدا می‌بایست متن‌های مورد نظر به فرمت استاندارد تبدیل شود لذا در روش پیشنهادی ابتدا یک پیش‌پردازش مانند ریشه‌یابی، یکسان‌سازی حروف و دیگر موارد توسط ابزارهای موجود صورت پذیرفت تا از مکالمات استاندارد در فرآیند مدل‌سازی استفاده گردد. همانطور که می‌دانیم هر چه قدر فرآیند استخراج ویژگی دقیق‌تر باشد امکان ارائه یک مدل مناسب‌تر، بیشتر فراهم می‌گردد. لذا ویژگی‌های متعددی پیشنهاد گردید که با استفاده از آن بتوان تمامی ابعاد در زمینه ارزیابی را در نظر گرفت. از آنجایی که هدف در این رساله ارائه یک روش آماری بود که بتواند مستقل از زبان بر روی هر سیستمی عمل نماید. تمامی ویژگی‌ها با در نظر گرفتن این پیش زمینه آماری پیشنهاد گردیدند. ۳۱ ویژگی مختلف در این رساله پیشنهاد گردید که این ویژگی‌ها در هیچ یک از کارهای ارزیابی قبلی مشاهده نگردیده بود و بعضاً ویژگی‌هایی که در زمینه‌های دیگر استفاده شده بودند متناسب با فرآیند ارزیابی سیستم‌های IQA به روز رسانی شدند. پس از استخراج ویژگی، ماتریس ویژگی با استفاده از مقادیر ویژگی‌ها و امتیاز داده شده توسط کاربران ایجاد شد. بر اساس بررسی صورت گرفته بر روی امتیازات داده شده توسط کاربران، امکان دسته‌بندی مکالمات به سه سطح خوب، متوسط و ضعیف فراهم شد که برای انجام این دسته‌بندی به صورت اتوماتیک از روش K-means استفاده گردید. لازم به ذکر است که مکالمات انگلیسی پایگاه داده نیز به صورت دستی به سه سطح خوب، متوسط و ضعیف تبدیل شدند برای انجام اینکار، به صورت تصادفی یکسری از خطوط مکالمه با خطوط دیگر مکالمه‌ها جابجا گردید که متناسب با تعداد خطوط جابجا شده در این سه گروه قرار گرفتند. در نهایت برای مدل‌سازی از روش رگرسیون استفاده گردید. معادلات متعدد رگرسیونی مورد آزمایش قرار گرفت تا بهترین پاسخ حاصل گردد. برای ارزیابی مدل رگرسیونی از سه معیار  $R^2$ ، RMSE و MAPE بهره گرفتیم که جزء معیارهای مرسوم در

ارزیابی مدل رگرسیونی می‌باشد. مدل‌های رگرسیونی در نظر گرفته شده هم شامل مدل‌های خطی و هم غیرخطی بود که هر کدام با در نظر گرفتن تعدادی ویژگی حاصل شده بود. در نهایت با توجه به کامل شدن فهرست ویژگی‌های در نظر گرفته برای بدست آوردن مدل رگرسیونی برای هر کلاس از برنامه‌نویسی بیان ژن استفاده گردید. در نهایت با تقسیم داده‌ها به مجموعه آزمون و آموزش به ترتیب هر کدام ۳۰ و ۷۰ درصد داده‌ها را تشکیل می‌دادند با آزمون مدل بدست آمده معیار خطا RMSE با درصد خطایی برابر ۰/۰۹ گزارش گردید. با توجه به اینکه در این معیار هر چه عدد حاصل نزدیک به صفر باشد مدل حاصل شده بهتر است نتیجه حاصل حاکی از مدل مناسب ارائه شده می‌باشد. این نکته لازم بذکر است که با توجه به نگاه متفاوت در حل مسئله مطرح شده جهت ارزیابی سیستم‌های IQA، امکان مقایسه نتایج با کار مشابه در این زمینه وجود نداشت و تنها امکان مقایسه امتیازات بدست آمده با نظرات واقعی داده شده توسط ارزیاب‌ها وجود داشت. نتایج حاصل از این مقایسه که در جدول ۴-۱۳ نمایش داده شده، حاکی از قابل قبول بودن مدل رگرسیونی ارائه شده می‌باشد.

## ۵-۲ کارهای آینده

با توجه به اینکه امکان دسترسی کامل به تمامی سیستم‌های IQA موجود فراهم نبود. در صورتیکه دسترسی به این سیستم‌ها فراهم گردد، این امکان وجود دارد تا بتوان با در نظر گرفتن ویژگی‌های بیشتری در ارزیابی و مدل‌سازی بهتری جهت حذف کامل ارزیاب صورت داد. یکی از کارهایی که می‌توان در زمینه ارزیابی این سیستم‌ها انجام داده استفاده از گزارش عملکرد (لاگ فایل سیستم‌ها) این سیستم‌ها در ارزیابی می‌باشد. در این روش می‌توان با استفاده از تکنیک‌های یادگیری ماشین به پردازش اتوماتیک لاگ فایل سیستم پرداخت و امتیاز کسب شده از این نوع عملیات را به عنوان یک ویژگی در نظر نهایی ارزیاب مدل نمود.

استفاده از تکنیک‌های یادگیری ماشین در تحلیل لاگ فایل‌ها خود یکی دیگر از کارهایی است که می‌تواند به روند اتوماتیک کردن این فرآیند کمک نموده تا بتوان مدل مناسب‌تری برای ارزیابی ارائه نمود. یک کار اولیه در این زمینه بر روی سیستم IQA موجود در آزمایشگاه صورت پذیرفت ولی به دلیل عدم دسترسی به فایل گزارش‌گیری (لاگ فایل) دیگر سیستم‌ها امکان تست و تعمیم به دیگر سیستم‌ها وجود نداشت. همچنین می‌توان با دسترسی به اطلاعات دیگر سیستم‌ها در مورد مدل‌سازی قابلیت اطمینان این سیستم‌ها نیز یک روش اتوماتیک ارائه نمود. یعنی از روش‌های مرسوم که در تحلیل نرم‌افزارها مانند معیار CK و MOOD وجود دارد، استفاده نمود. یکی دیگر از روش‌هایی که می‌توان در این زمینه از آن بهره گرفت، استفاده از یادگیری عمیق در مدل‌سازی نظرات مربوط با ارزیاب‌ها می‌باشد. البته در این رساله از این روش تنها در پیش‌بینی امتیاز بر اساس ویژگی‌های کسب شده استفاده گردید که می‌توان در آینده به شرط فراهم شدن حجم مکالمات مناسب مدل‌سازی نیز در این زمینه با استفاده از یادگیری عمیق انجام داد. البته لازم به ذکر است که در شبکه‌های یادگیری عمیق نیازی به استخراج ویژگی نمی‌باشد و تنها کافی است با داشتن مکالمات و تبدیل آن‌ها به بردار با استفاده از روش word2vec بتوان شبکه عمیقی ایجاد نمود که مدل مناسب و دقیق‌تری را در اختیار ما قرار دهد. از طرفی می‌توان از روش‌های دیگر مانند GMM<sup>1</sup> برای بدست آوردن مدل هر کلاس نیز استفاده نمود. در این راه باید گشتاورهای مختلف را مورد آزمون قرار داد تا مدل مطلوب حاصل گردد. همچنین پیشنهاد می‌گردد که در آینده و با در نظر گرفتن اینکه هر کدام از این ویژگی‌ها چه ابعادی از نظرات انسان‌ها را پوشش می‌دهد، روشی برای اثبات پوشش این ابعاد (مانند آزمایش تعیین انسجام) معرفی کرد.

---

<sup>1</sup> Gaussian Mixture Models

## واژه نامه

<b>C</b>		Mean Absolute Percentage Deviation	میانگین درصد انحراف
Community Question Answering	پرسش و پاسخ اجتماعی	Mean Reciprocal Rank	میانگین معکوس رتبه
Confidence weight score	امتیاز وزن اطمینان	Mean Absolute Error	میانگین قدرمطلق خطا
Cumulative Density Function	تابع توزیع تجمعی	<b>N</b>	
<b>F</b>		N-gram	ان - گرم
False Positive	مثبت کاذب	Natural Language Processing	پردازش زبان طبیعی
False Negative	منفی کاذب	<b>Q</b>	
Follow-up Questions	سوالات متوالی	Question Answering system	سیستم پرسش و پاسخ
<b>G</b>		<b>R</b>	
Gene Expression Programming	برنامه نویسی بیان ژن	Root Mean Square Error	میانگین مجموع مربعات خطا
<b>I</b>		Regression	رگرسیون
Interactive Question Answering system	سیستم پرسش و پاسخ تعاملی	Recurrent Neural Network	شبکه عصبی بازگشتی
<b>K</b>		<b>S</b>	
K-means	k- میانگین	Skip-gram	ان - گرم
<b>L</b>		Support Vector Machine	ماشین بردار پشتیبان
Long Short Term Memory	حافظه طولانی کوتاه مدت	<b>T</b>	
<b>M</b>		Text Retrieval Conference	کنفرانس بازیابی متون
Mean Absolute Percentage Error	میانگین درصد قدرمطلق خطا	True Positive	مثبت درست



## منابع:

- [۱] شهرآیینی. س. (۱۳۹۴)، پایان نامه ارشد: "سیستم پاسخگویی تعاملی با استفاده از تکنیک های هوش مصنوعی"، مرکز آموزش های الکترونیکی دانشگاه صنعتی شاهرود.
- [2] Yu, J., Qiu, M., Jiang, J., Huang, J., Song, S., Chu, W., & Chen, H. (2018), "Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce", In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 682-690.
- [3] S. K. Dwivedi, and V. Singh, (2013), "Research and reviews in question answering system", *Journal of Procedia Technology.*, vol. 10, pp. 417-424.
- [4] A. Mishra, and S. K. Jain, (2016), "A survey on question answering systems with classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 28, no. 3, pp. 345-361.
- [5] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, (2015) "Question answering systems: survey and trends," *Procedia Computer Science*, vol. 73, pp. 366-375.
- [6] B. F. Green Jr, A. K. Wolf, C. Chomsky, and K. Laughery, (1961 ) "Baseball: an automatic question-answerer," *IRE-AIEE-ACM computer conference*, pp. 219-224.
- [7] W. A. Woods, (1973) "Progress in natural language understanding: an application to lunar geology," *national computer conference and exposition*, pp. 441-450.
- [8] Lee, S., Noh, H., Lee, J., Lee, K., Lee, G.G., Sagong, S. and Kim, M., (2011), "On the effectiveness of robot-assisted language learning", *Journal of ReCALL*, vol. 23, no.1, pp.25-58.
- [9] J Hirschberg, Julia, and Christopher D. Manning, (2015), "Advances in natural language processing." *Journal of Science*, vol. 349, no. 6245, pp. 261-266.
- [10] Bao, J., Duan, N., Zhou, M., & Zhao, T. (2014), "Knowledge-based question answering as machine translation", In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol.1, pp. 967-976.
- [11] E. Riloff, and M. Thelen, (2000), "A rule-based question answering system for reading comprehension tests," *Proceedings of the ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding systems*, vol. 8, no.6, pp. 13-19.
- [12] X. Hao, X. Chang, and K. Liu, (2007), "A Rule-based Chinese Question Answering System for Reading Comprehension Tests," *Intelligent Information Hiding and Multimedia Signal Processing, Third International Conference.*, vol. 2, pp. 325-329.

- [۱۳] حجازی. م. میریان حسین آبادی، آ. نشاطی. ک. افقی. ب. و درودی. ا. (۱۳۸۳) "سیستم پرسش و پاسخ مبتنی بر هستان شناسی برای حوزه مخابرات با قابلیت استخراج و دسته بندی خودکار مستندات"، مجله علوم و مهندسی کامپیوتر، شماره ۱-۳ (الف)، دوره ۲، ص ۱-۱۷.
- [۱۴] یوسفی. م. (۱۳۹۱)، "طراحی و پیاده سازی سیستم پرسش و پاسخ در یک دامنه محدود"، دانشکده فنی و مهندسی، موسسه آموزش عالی غیرانتفاعی و غیر دولتی نبی اکرم تبریز.
- [۱۵] خانی. ا. و. ساجدی. ه. (۱۳۹۵) "جويا: یک سیستم پرسش و پاسخ فارسی"، مجله علمی- ترویجی علوم رایانش، شماره ۳، دوره ۱۱، ص ۵۱-۶۶.
- [16] Kalyanpur, A., Murdock, J. W. (2015), "Unsupervised entity-relation analysis in IBM watson", In *Proceedings of the Third Annual Conference on Advances in Cognitive Systems ACS*, pp. 1-12.
- [17] A. Moschitti, (2003), "Answer filtering via text categorization in question answering systems", In *Tools with Artificial Intelligence, Proceedings 15th IEEE International Conference*, pp. 241-248.
- [18] K. Zhang, and J. Zhao, (2010), "A Chinese question-answering system with question classification and answer clustering," *Fuzzy Systems and Knowledge Discovery (FSKD), Seventh International Conference on*, vol. 6, pp. 2692-2696.
- [19] S. Quarteroni, and S. Manandhar, (2009), "Designing an interactive open-domain question answering system," *Natural Language Engineering*, vol. 15, no. 1, pp. 73-95.
- [20] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal, (2001), "Bridging the lexical chasm: statistical approaches to answer-finding", In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 192-199.
- [21] D. Cai, Y. Dong, D. Lv, G. Zhang, and X. Miao, (2006), "A Web-based Chinese question answering with answering validation", *Natural Language Processing and Knowledge Engineering, IEEE NLP-KE'05. Proceedings of IEEE International Conference on*, pp. 499-502.
- [22] T. Gunawardena, N. Pathirana, M. Lokuhetti, R. Ragel, and S. Deegalla, (2015), "Performance Evaluation Techniques for an Automatic Question Answering System", *International Journal of Machine Learning and Computing*, vol. 5, no. 4, pp. 294.
- [23] B. W. van Schooten, R. Den Akker, S. Rosset, O. Galibert, A. Max, and G. Illouz, (2009), "Follow-up question handling in the IMIX and Ritel systems: A comparative study", *Natural Language Engineering*, vol. 15, no. 1, pp. 97-118.
- [24] Y. Qu, and N. Green, (2004), "A constraint-based approach for cooperative information-seeking dialogue", *Proceedings of the International Natural Language Generation Conference*, pp. 136-143.

- [25] S. Vargas, F. Weng, and H. Pon-Barry, (2011), "Interactive question answering and constraint relaxation in spoken dialogue systems", *Natural Language Engineering*, vol. 15, no. 1, pp. 9-30.
- [26] B. Magnini, M. Speranza, and V. Kumar, (2009), "Towards interactive question answering: An ontology-based approach", *Semantic Computing, ICSC'09. IEEE International Conference*. pp. 612-617.
- [27] I. Dornescu, and C. Orasan, (2010), "Interactive QA using the QALLME framework" *International Journal of Computational Linguistics and Applications*, vol. 1, no. 1-2, pp. 233-247.
- [28] L. Liu, Q. Qi, and F. Li, (2010), "Ontology-based interactive question and answering system", *Internet Technology and Applications, International Conference*, pp. 1-4.
- [29] Luo, G. and Tang, C., (2008), "On iterative intelligent medical search", *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3-10.
- [30] Hovy, E., Oberlander, J. and Reithinger, N., (2011), "IMIX: Good Questions, Promising Answers". *In Interactive Multi-modal Question-Answering, Book chapter, Springer*, pp. 271-279.
- [31] O. Galibert, G. Illouz, and S. Rosset, (2005), "Ritel: an open-domain, human-computer dialog system", *Ninth European Conference on Speech Communication and Technology*, pp. 909-913.
- [32] R. Bernardi, and M. Kirschner, (2010), "Context modeling for iqa: the role of tasks and entities," *Proceedings of the workshop on Knowledge and Reasoning for Answering Questions*, pp. 25-32.
- [33] M. Kirschner, and R. Bernardi, (2010), "Exploring topic continuation follow-up questions using machine learning", *Proceedings of Human Language Technologies, the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, pp. 13-18.
- [34] M. Kirschner, R. Bernardi, M. Baroni, and L. T. Dinh, (2011), "Analyzing interactive QA dialogues using logistic regression models", *Congress of the Italian Association for Artificial Intelligence*, pp. 334-344.
- [35] B. Van Schooten, and R. op den Akker, (2012), "Vidiam: Corpus-based development of a dialogue manager for multimodal question answering," *Interactive Multi-modal Question-Answering, Book chapter Springer*, pp. 25-56.
- [36] D. Wang, (2011), "Answering contextual questions based on ontologies and question templates", *Frontiers of Computer Science in China*, vol. 5, no. 4, pp. 405-418.

- [37] U. Waltinger, A. Breuing, and I. Wachsmuth, (2012) "Connecting question answering and conversational agents", *KI-Künstliche Intelligenz*, vol. 26, no. 4, pp. 381-390.
- [38] J. Tague-Sutcliffe, (1992), "The pragmatics of information retrieval experimentation, revisited", *Information Processing & Management*, vol. 28, no. 4, pp. 467-490.
- [39] O. Kolomiyets, and M.-F. Moens, (2011), " A survey on question answering technology from an information retrieval perspective," *Information Sciences*, vol. 181, no. 24, pp. 5412-5434.
- [40] Buscaldi, D., Rosso, P., Gómez-Soriano, J. M., & Sanchis, E. (2010), "Answering questions with an n-gram based passage retrieval engine", *Journal of Intelligent Information Systems*, vol. 34, no.2, pp. 113-134.
- [41] M. Spitters, M. De Boni, J. Zavrel, and R. Bonnema, (2009), "Learning effective and engaging strategies for advice-giving human-machine dialogue", *Natural Language Engineering*, vol. 15, no. 3, pp.355-378.
- [42] S. Harabagiu, A. Hickl, J. Lehmann, and D. Moldovan, (2005) "Experiments with interactive question-answering", *Proceedings of the 43rd annual meeting on Association for Computational Linguistics*, pp. 205-214.
- [43] D. Kelly, P. B. Kantor, E. L. Morse, J. Scholtz, and Y. Sun, (2009), "Questionnaires for eliciting evaluation data from users of interactive question answering systems", *Natural Language Engineering*, vol. 15, no. 1, pp. 119-141.
- [44] D. Kelly, P. B. Kantor, E. L. Morse, J. Scholtz, and Y. Sun, (2006), "User-centered evaluation of interactive question answering systems", *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL*, pp. 49-56.
- [45] Y. Sun, P. B. Kantor, and E. L. Morse, (2011), "Using cross-evaluation to evaluate interactive QA systems", *Journal of the Association for Information Science and Technology*, vol. 62, no. 9, pp. 1653-1665.
- [46] S. Quarteroni, (2007), PhD.thesis, "Advanced techniques for personalized, interactive question answering", computer.depart, University of York.
- [47] C. Shah, and J. Pomerantz, (2010), "Evaluating and predicting answer quality in community QA", *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 411-418.
- [48] Wacholder, N. (2011), "Interactive query formulation", *Annual review of information science and technology*, vol. 45, no.1, pp. 157-196.
- [49] P. Forner, D. Giampiccolo, B. Magnini, A. Peñas, Á. Rodrigo, and R. Sutcliffe, (2010), "Evaluating multilingual question answering systems at CLEF", *Multilingual and Multimodal Information Access Evaluation*, pp.2774-2781.
- [50] Rodrigo, A., Peñas, A, (2017), "A study about the future evaluation of Question-Answering systems", *Journal of Knowledge-Based Systems*, vol.137, pp.83-93.

- [51] M. Sanderson, (2010), "Test collection based evaluation of information retrieval systems", *Foundations and Trends® in Information Retrieval*, vol. 4, no.4, pp.247-375.
- [52] N. Konstantinova, C. Orasan (2013), "Interactive question answering", In *Emerging Applications of Natural Language Processing: Concepts and New Research*, IGI Global, pp. 149-169.
- [53] T. Strzalkowski, and S. Harabagiu, (2006), "*Advances in open domain question answering*", *Springer Science & Business Media*, vol.33, no.4.
- [54] C. Guinaudeau, and M. Strube, (2013), "Graph-based local coherence modeling", *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 93-103.
- [55] C. Ferreira, (2002) , "Gene expression programming in problem solving", *Soft computing and industry*, Book chapter, Springer, pp. 635-653.
- [56] Karaboga, Dervis, Celal Ozturk, Nurhan Karaboga, and Beyza Gorkemli. "Artificial bee colony programming for symbolic regression", *Information Sciences*, vol. 209, pp. 1-15.
- [57] J. Brownlee, (2011), "Clever algorithms: nature-inspired programming recipes", *Creative Commons*, Melbourne, Australia, p.454.
- [58] Wang, D., & Nyberg, E. (2015), "A long short-term memory model for answer sentence selection in question answering", In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 2, pp. 707-712.



پیوست الف

رگرسیون

## رگرسیون خطی:

تحلیل‌ها و پیش‌فرض‌های گوناگونی در رابطه با رگرسیون وجود دارد که از جمله پیش‌فرض‌های مدل‌های رگرسیون می‌توان به موارد زیر اشاره نمود:

الف- وجود مدل نظری که احتمال وجود رابطه بین متغیرها و مدل را بالا ببرد.

ب- نرمال بودن توزیع.

ج- مساوی بودن واریانس‌ها بین سطوح متغیر وابسته و مستقل.

د- باید بین متغیر وابسته و مستقل مرزی وجود داشته باشد، از دو مشخصه جدا باشند. به بیان بهتر، نباید اشتراکی بین متغیر مستقل و وابسته وجود داشته باشد.

در علم آمار دو بخش اصلی برای رگرسیون وجود دارد: پارامتری و ناپارامتری. در رگرسیون پارامتری نوع ارتباط بین متغیرهای وابسته و مستقل شناخته شده است، اما ممکن است پارامترها مقادیری را شامل شوند که ناشناخته بوده و صلاحیت برآورد مجموعه داده‌ها را نداشته باشند. علائم ویژه رگرسیون ناپارامتری زمانی مشاهده می‌شود که آگاهی قبلی در مورد نوع واقعی تابعی که قرار است برآورد شود، وجود ندارد. در رگرسیون پارامتری نوعاً یک تعداد کم از پارامترها، که اغلب آن‌ها تفسیر فیزیکی (طبیعی) دارند، وجود دارد. به عبارت دیگر هدف اصلی رگرسیون برآورد مقادیر پارامتری باشد چرا که مفهوم اصلی آن‌ها می‌باشد. در مطالعه روش‌های تحلیل رگرسیون پنج دسته قابل ذکر می‌باشند که عبارتند از:

الف- رگرسیون خطی ساده و رگرسیون خطی چند گانه<sup>۱</sup>:

<sup>۱</sup> Simple&Multiple linear regression



رگرسیون ساده خطی یک ابزار آماری است که در آن به بررسی رابطه یک متغیر مستقل (پیش بین) و یک متغیر وابسته پرداخته می‌شود. حال اگر تعداد متغیرهای مستقل در این رابطه خطی بیش از یک عدد شود، مدل رگرسیون، خطی چندگانه نامیده می‌شود. معادله رگرسیون خطی ساده به شکل  $Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$  و رگرسیون خطی چندگانه به صورت  $Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$  می‌باشد.

ب- رگرسیون لجستیک دوگانه و رگرسیون لجستیک چندگانه:

رگرسیون لجستیک، هم چون رگرسیون خطی می‌باشد با این تفاوت که در این نوع پیش بینی، نتیجه نهایی در مورد سطح یا سطوح متغیر وابسته دو جوابی می‌باشد. به عبارت دیگر، در این رگرسیون متغیرهای مستقل در قالب مدلی ارائه می‌شوند و سعی در پیش بینی متغیر وابسته و یا سطوح آن در قالب یک حالتی از آری یا نه به عمل می‌آید. رگرسیون لجستیک دوگانه به حالتی اطلاق می‌شود که متغیر وابسته یک سطح دارد. رگرسیون لجستیک چندگانه به وضعیتی گفته می‌شود که متغیر وابسته بیش از یک سطح دارد. به طور کلی معادله رگرسیون لجستیک به شکل زیر می‌باشد.

$$p = \frac{e^{\beta_0 + \beta_1 \times \text{predictor}}}{1 + e^{\beta_0 + \beta_1 \times \text{predictor}}} \quad (1-1)$$

در این مدل،  $p$  احتمال این است که جواب ما به سؤال دوگانه آری شود. هر چه این احتمال به یک نزدیک شود، احتمال تحقق ویژگی در مورد سطح یا سطوح متغیر وابسته بالاتر می‌رود. در انتخاب مدل در رگرسیون چندگانه باید یکسری ملاحظات را در نظر گرفت. در رگرسیون چندگانه، چند متغیر مستقل مدلی را می‌سازند که مقدار متغیر وابسته را پیش بینی می‌کنند. از آنجا که امکان دارد این متغیرهای مستقل با یکدیگر رابطه داشته باشند، همیشه این نگرانی وجود دارد که آیا مدل به دست آمده اعتبار دارد یا خیر. برای حل این مشکل، روش‌هایی برای وارد نمودن متغیرهای مستقل درون مدل رگرسیون پیشنهاد شده‌اند:

- انتخاب رو به جلو<sup>۱</sup>: در این روش به ترتیب تک تک متغیرهای مستقل وارد مدل رگرسیون می‌شوند و در هنگام اضافه شدن، یک معیار برای پذیرش متغیر مستقل در نظر گرفته می‌شود.
- حذف پسرو<sup>۲</sup>: در این روش به صورت تصادفی و ابتکاری همه متغیرهای مستقل وارد مدل می‌شوند و به صورت مرحله به مرحله از یک نوع معیار خاص برای حذف هر کدام از متغیرها بهره جسته می‌شود. در صورتی که با توجه به معیار مورد نظر وجود متغیر غیر ضروری باشد، آن متغیر از مدل حذف می‌شود.
- نشانگر گام به گام<sup>۳</sup>: در این روش هم چون انتخاب رو به جلو متغیرها تک تک اضافه می‌شوند با این تفاوت که بعد از اضافه شدن، تک تک متغیرها بر اساس معیار روش حذف پس رو بررسی می‌شوند.

در مورد استفاده از رگرسیون لجستیک باید به موارد زیر توجه داشت:

الف- این نوع از رگرسیون به طور گسترده ای برای مشکلات طبقه بندی مورد استفاده قرار می‌گیرد.

ب- در رگرسیون لجستیک نیازی به رابطه خطی بین متغیر وابسته و مستقل وجود نیست. که این عامل می‌تواند انواع مختلفی از روابط را پشتیبانی نماید، زیرا یک تغییر  $\log$  غیر خطی را به نسبت احتمالات پیش‌بینی شده اعمال می‌کند.

ج- برای جلوگیری از بیش برآزش<sup>۴</sup> و کم برآزش<sup>۵</sup>، باید همه متغیرهای مهم را شامل شود. یک رویکرد خوب برای اطمینان از این عمل، استفاده از یک روش گام به گام برای تخمین رگرسیون

<sup>1</sup> Forward Selection

<sup>2</sup> Backward Elimination

<sup>3</sup> Stepwise Entry

<sup>4</sup> Over fitting

<sup>5</sup> Under fitting

لجستیک است که این به اندازه‌های نمونه بزرگ نیاز دارد، چون برآورد حداکثر احتمال در نمونه‌های کم، کم‌تر از حداقل مربعات است. متغیرهای مستقل نباید با یکدیگر همبستگی داشته باشند، به عنوان مثال هم خطی بین متغیرهای مستقل. با این حال، ما گزینه‌هایی داریم تا اثرات تعاملی متغیرهای مطلق در تحلیل و مدل را در بر بگیریم.

د- اگر مقادیر متغیر وابسته ترتیبی باشد، آن را رگرسیون لجستیک عادی می‌نامند.

ه- اگر متغیر وابسته چند کلاس باشد، آن را به عنوان رگرسیون لجستیک چندجمله‌ای شناخته می‌شود

پ- رگرسیون به روش تخمین منحنی<sup>۱</sup>:

همانند دیگر تحلیل‌های رگرسیونی، قبل از انجام این تحلیل می‌بایست در مورد داده‌های پژوهش و این که به چه مدلی نزدیک‌تر می‌باشند، شناخت اولیه داشته باشد با استفاده از این قابلیت داده‌ها و نمودارهای مرتبط با آنان با استفاده از ۱۱ مدل آزمون می‌شوند. این یازده مدل آزمون به شرح زیر می‌باشند:

$$y = b_0 + (b_1 \times t) \quad \text{۱- رگرسیون خطی}^2$$

$$y = b_0 + (b_1 \times \ln(t)) \quad \text{۲- رگرسیون لگاریتمی}^3$$

$$y = b_0 + \left(\frac{b_1}{t}\right) \quad \text{۳- رگرسیون معکوس}^4$$

$$y = b_0 + (b_1 \times t) + (b_2 \times t^2) \quad \text{۴- رگرسیون درجه دو}^5$$

<sup>1</sup> Curve Estimation

<sup>2</sup> Linear

<sup>3</sup> Logarithmic

<sup>4</sup> Inverse

<sup>5</sup> Quadratic

$$y = b_0 + (b_1 \times t) + (b_2 \times t^2) + (b_3 \times t^3) \quad \text{۵- رگرسیون درجه سه}^1$$

$$\ln(y) = \ln(b_0) + (b_1 \times \ln(t)) \quad \text{۶- رگرسیون پاور}^2$$

$$\ln(y) = \ln(b_0) + (\ln(b_1) \times t) \quad \text{۷- رگرسیون مرکب}^3$$

$$y = e^{b_0 + (\frac{b_1}{t})} \quad \text{۸- رگرسیون منحنی}^4$$

$$y = \frac{1}{(\frac{1}{u} + (b_0 \times (b_1 \times t)))} \quad \text{۹- رگرسیون لجستیک}^5$$

$$y = e^{b_0 + b_1 \times t} \quad \text{۱۰- رگرسیون رشد}^6$$

$$y = b_0 \times e^{b_1} \quad \text{۱۱- رگرسیون نمایی}^7$$

۵- رگرسیون رتبه‌ای

در رگرسیون رتبه‌ای اثر متغیر و یا متغیرهای مستقل بر روی متغیر وابسته‌ای که سطوح مختلف و رتبه‌ای دارد، پیش‌بینی می‌شود.

ت- رگرسیون پروبیت (تحلیل پروبیت)<sup>۸</sup>

هنگامی که خروجی (متغیر وابسته) دو حالت داشته باشد (آری یا خیر) و هدف بررسی شدت متغیر مستقل در پیش‌بینی این دو حالت باشد، از تحلیل پروبیت استفاده می‌شود. رگرسیون ناپارامتری معمولاً در فرضیات خطی آزاد می‌باشد و ما را به شرح داده‌های بصری، ساختار غیرپوششی در داده‌ها که ممکن است به نحوی گمشده باشد، قادر می‌سازد. بنابراین خیلی از روش‌های رگرسیون ناپارامتری

<sup>1</sup> Cubic

<sup>2</sup> Power

<sup>3</sup> Compound

<sup>4</sup> S-curve

<sup>5</sup> Logistic

<sup>6</sup> Growth

<sup>7</sup> Exponential

<sup>8</sup> Probit analysis

هنگامی که تعداد متغیرهای مستقل در مدل زیاد می‌باشد به خوبی اجرا نمی‌شوند. پراکندگی داده‌ها در این مجموعه سبب می‌شود بر آوردهای واریانس به اندازه غیر قابل پذیرش بزرگ شود، مگر آنکه حجم نمونه فوق العاده بزرگ باشد. قابلیت تفسیر یکی دیگر از مسائل رگرسیون ناپارامتری است که بر پایه کرنل و هموارسازی برآوردهای خط می‌باشد. معدل گیری موضعی برآوردهای کرنل، رگرسیون ناپارامتری نیرومند، رگرسیون و هموارسازی دسته‌های باریک، استنتاج آماری برای رگرسیون ناپارامتری در تجزیه و تحلیل داده‌ها، رگرسیون چند متغیر ناپارامتری به انضمام مدل‌های رگرسیون افزایشی، رگرسیون ناپارامتری تعمیم یافته و مدل‌های تعمیم یافته افزایشی همگی جز رگرسیون ناپارامتری می‌باشند. رگرسیون ناپارامتری فرضیات کمینه در مورد وابستگی میانگین  $Y$  بر روی  $X$ ‌ها را درست می‌کند. این جریان کوتاه برآوردهای رگرسیون ناپارامتری را به دو صورت برای تحلیل رگرسیون ساده (یک  $X$  تنها) - موسوم به نمودار پراکندگی هموارساز - و تحلیل رگرسیون چند متغیره (چندین  $X$ ) معرفی می‌کند. چند ملاحظه نیز برای روش‌های استنتاج آماری برای رگرسیون ناپارامتری وجود دارد، که شبیه بکار گرفته شده برای حداقل مربعات خطی می‌باشد.

ث- رگرسیون چند جمله‌ای:

اگر درجه توان یک متغیر در معادله رگرسیون بیشتر از یک باشد آنگاه آن یک معادله رگرسیونی چند جمله‌ای خواهد بود. در این حالت، مناسب‌ترین خط یک خط مستقیم نمی‌باشد بلکه یک منحنی خواهد بود که متناسب با نقاط داده‌ها خواهد بود. به نکته‌ای که باید در زمان استفاده از این معادله رگرسیونی توجه داشت این است که در حالی که ممکن است وسوسه استفاده از چند جمله‌ای درجه بالاتر برای رسیدن به خطای کمتری وجود داشته باشد، این می‌تواند منجر به بیش‌برازش شود. همیشه روابط را ترسیم کنید تا تناسب را ببینید و بر روی اطمینان از اینکه این منحنی با ماهیت مشکل متناسب است، تمرکز کنید. به خصوص برای منحنی به سمت انتهای آن‌ها نگاه کنید و ببینید آیا این

شکل‌ها و گرایش‌ها منطقی هستند یا خیر؟ چند جمله‌ای‌های بالاتر می‌توانند منجر به تولید نتایج عجیب درباره برون‌یابی شوند.

### ج- رگرسیون مرحله‌ای<sup>۱</sup>:

این شکل از رگرسیون زمانی استفاده می‌شود که ما با چندین متغیر مستقل سروکار داریم. در این روش، انتخاب متغیرهای مستقل با کمک یک فرآیند خودکار انجام می‌شود که شامل هیچ مداخله انسانی نمی‌شود. این شاهکار با مشاهده مقادیر آماری مانند  $t$ -test،  $R$ -square و معیار AIC برای تشخیص متغیرهای معنی‌دار حاصل می‌شود. رگرسیون Stepwise اساساً متناسب با مدل رگرسیون با اضافه کردن / حذف یک متغیر براساس یک معیار مشخص است. برخی از رایج‌ترین موارد استفاده این روش‌های رگرسیون در زیر لیست شده‌اند:

- رگرسیون مرحله‌ای استاندارد دو چیز را انجام می‌دهد. این رگرسیون ابتدا متغیری را اضافه می‌کند و پیش‌بینی‌ها را که در هر مرحله مورد نیاز است را حذف می‌کند.
- انتخاب پیش‌رو<sup>۲</sup> با مهم‌ترین پیش‌بینی‌کننده در مدل شروع می‌شود و متغیر را برای هر مرحله اضافه می‌کند.
- حذف بعد با تمام پیش‌بینی‌ها در مدل شروع می‌شود و کم‌ترین متغیر را برای هر مرحله از بین می‌برد.

هدف از این روش مدل‌سازی، به حداکثر رساندن توان پیش‌بینی با کم‌ترین تعداد متغیرهای پیش‌بینی‌کننده است. این یکی از روش‌ها برای رسیدگی به ابعاد بالاتر مجموعه داده‌ها است.

### ج- رگرسیون ریج<sup>۳</sup>:

<sup>1</sup> Stepwise Regression

<sup>2</sup> Forward selection

<sup>3</sup> Ridge regression

رگرسیون ريج در زمانی مورد استفاده قرار می‌گیرد که داده‌ها از متغیرهای مستقل چندگانه رنج می‌برند به عبارتی دیگر، متغیرهای مستقل به شدت به یکدیگر هم‌بسته هستند. در چند خطی چندگانه، حتی اگر کم‌ترین مربعات (OLS<sup>1</sup>) بی طرفانه باشند، واریانس آن‌ها بزرگ است که مقدار مشاهده شده دور از مقدار حقیقی را منحرف می‌کند. با افزودن درجه‌ای از مقدار اولیه بایاس به برآورد رگرسیون، رگرسیون ريج خطاهای استاندارد را کاهش می‌دهد. در یک معادله خطی، خطاهای پیش‌بینی می‌توانند به دو جز فرعی تجزیه شوند. اول ناشی از انحراف و دوم ناشی از واریانس است. خطای پیش‌بینی می‌تواند ناشی از هر کدام از این دو یا هر دو جز باشد. رگرسیون ريج مشکل چند خطی را از طریق پارامتر انقباض و  $\lambda$  حل می‌کند. به معادله زیر نگاه کنید.

$$y = \arg \min_{\beta \in \mathbb{R}^p} \underbrace{\|y - X\beta\|_2^2}_{Loss} + \lambda \underbrace{\|\beta\|_2^2}_{Penalty} \quad (۲-۱)$$

در این معادله، ما دو جز داریم. اول عبارت مربع است و دیگری  $\lambda$  از جمع  $\beta_2$  (مربع بتا) است که در آن  $\beta$  ضریب است. این مقدار به حداقل مربع اضافه می‌شود تا پارامتر را کوچک کند که واریانس بسیار پایینی داشته باشد. نکاتی که در مورد استفاده از این رگرسیون باید توجه نمود به شرح زیر است:

- فرضیات این رگرسیون مانند رگرسیون حداقل مربعات معمولی هستند، به جز نرمال بودن که فرض نمی‌شود.
- این روش ارزش ضرایب را کاهش می‌دهد اما به صفر نمی‌رسد، که هیچ ویژگی انتخاب ویژگی را نشان نمی‌دهد.
- یک روش پایدارسازی است و از روش‌های خودکار استفاده می‌کند.

<sup>1</sup> Ordinary least square elimination

ح- رگرسیون لسو<sup>۱</sup>:

همانند رگرسیون ریج، رگرسیون لسو<sup>۲</sup> (که مخفف کمترین انقباض مطلق و اپراتور انتخاب است) اندازه مطلق ضرایب رگرسیون را مورد بررسی قرار می‌دهد. علاوه بر این، قادر به کاهش تغییرپذیری و بهبود دقت مدل‌های رگرسیون خطی نیز است. به معادله زیر نگاه کنید:

$$y = \arg \min_{\beta \in \mathbb{R}^p} \underbrace{\|y - X\beta\|_2^2}_{Loss} + \lambda \underbrace{\|\beta\|_1}_{Penalty} \quad (3-1)$$

رگرسیون لسو از رگرسیون ریج به گونه‌ای متفاوت است که از مقادیر مطلق در تابع جریمه به جای مربع استفاده می‌کند. این امر منجر به penalizing (یا محدود کردن مجموع مقادیر مطلق برآوردها) می‌شود که باعث می‌شود برخی از تخمین‌های پارامتری دقیقاً صفر شوند. پس از اعمال مجازات، برآوردهای بیشتر به صفر مطلق تبدیل می‌شوند. این امر منجر به انتخاب متغیر از  $n$  متغیر می‌شود. نکات مهم در مورد این رگرسیون عبارتند از:

- فرض این رگرسیون مانند رگرسیون خطی حداقل مربعات است، به جز حالت نرمال بودن که نمی‌توان فرض کرد.
- این رگرسیون باعث می‌شود که ضریب صفر باشد (دقیقا صفر)، که قطعا به انتخاب ویژگی کمک می‌کند.
- این یک روش پایدارسازی است و از تنظیمات  $L1$  استفاده می‌کند.
- اگر گروه پیش‌بینی‌کننده‌ها بسیار همبسته باشند، lasso تنها یکی از آن‌ها را انتخاب کرده و بقیه را به صفر کاهش می‌دهد.

<sup>۱</sup> Lasso regression

<sup>۲</sup> Least Absolute Shrinkage and Selection Operator (lasso)

<sup>۳</sup> Regularization



## خ- رگرسیون Elasticnet:

- ElasticNet ترکیبی از Lasso و تکنیک های رگرسیون ریج است. با  $L_1$  و  $L_2$  قبل از آن به عنوان یک سیستم تنظیم کننده آموزش داده شده است. Net-Elastic هنگامی مفید است که ویژگی های متعددی وجود دارد که همبستگی دارند. به احتمال زیاد Lasso یکی از این ها را تصادفی انتخاب می کند، در حالی که Net-Elastic می تواند هر دو را انتخاب کند.

$$y = \arg \min_{\beta \in \mathbb{R}^p} (\|y - X\beta\|^2 + \lambda \|\beta\|_1 + \lambda_2 \|\beta\|^2) \quad (4-1)$$

یک مزیت این نوع رگرسیون در این است که به Elastic اجازه می دهد تا برخی از ثبات ریج را در زمان چرخش به ارث ببرد. نکات مهم در زمان استفاده از این رگرسیون عبارتند از:

- این گروه تاثیر گروه را در مورد متغیرهای بسیار همبسته مورد تشویق قرار می دهد .
- هیچ محدودیتی بر تعداد متغیرهای انتخاب شده وجود ندارد
- می تواند از انقباض دوگانه رنج ببرد



# پيوسٽ ب

## برنامه نویسی بیان ژن

## برنامه نویسی بیان ژن:

برنامه نویسی بیان ژن تعمیم یافته الگوریتم ژنتیک<sup>۱</sup> است که بر اساس تئوری داروین ارائه شده و توسط فریرا در سال ۱۹۹۹ ابداع شد [۵۵]. GEP نیز همانند الگوریتم ژنتیک و برنامه ریزی ژنتیک<sup>۲</sup> یک الگوریتم ژنتیکی است که از جمعیتی از افراد استفاده کرده و آنها را مطابق برازندگی انتخاب می‌کند و تغییرات ژنتیکی را با استفاده از یک یا چند عملگر ژنتیکی اعمال می‌نماید. تفاوت اساسی بین این سه الگوریتم، مربوط به ماهیت افراد آنهاست. به طوری که در GA افراد رشته‌های خطی با طول ثابت (کروموزوم) و در GP نهادهای غیرخطی با اندازه‌ها و اشکال متفاوت (درختان تجزیه) می‌باشند. در حالی که در GEP افراد به صورت رشته‌های خطی با طول ثابت (ژنوم یا کروموزوم) کدگذاری شده (مشابه با آن چه که در الگوریتم ژنتیک استفاده می‌شود) و سپس به شکل نهادهای غیر خطی با اندازه‌ها و اشکال متفاوت (یعنی نمایش دیاگرام ساده یا بیان درختی مشابه با درختان تجزیه در برنامه ریزی ژنتیک) اظهار می‌شوند. برنامه ریزی بیان ژن یک تکنیک برنامه ریزی خودکار می‌باشد که راه حل مسئله را با استفاده از برنامه کامپیوتری ارائه می‌کند. در این روش در ابتدای فرآیند هیچگونه رابطه تابعی در نظر گرفته نشده و این روش قادر به بهینه سازی ساختار مدل و مؤلفه‌های آن می‌باشد. تحقیقات گذشته نشان می‌دهد که در استفاده از محاسبات نرم (مثل برنامه ریزی ژنتیک و شبکه‌های عصبی و) به بعضی از مشکلات مدل‌های فیزیکی غلبه شده است. قبل از مراحل اجرایی برنامه ریزی بیان ژن گام‌های مقدماتی زیر باید توسط کاربر تعیین شوند:

۱- مجموعه ترمینال‌ها (متغیرهای مسئله، اعداد ثابت تصادفی)

۲- مجموعه عملگرهای ریاضی مورد استفاده در فرمول‌ها

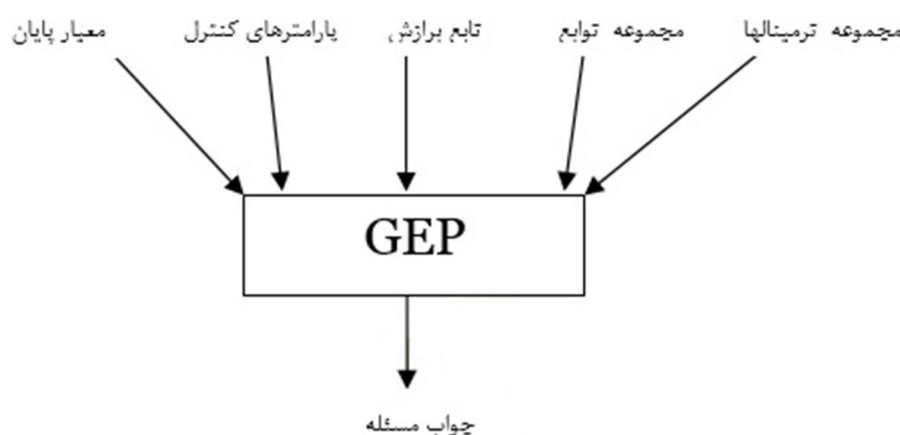
۳- انتخاب تابع برازش مناسب برای سنجش برازش فرمول‌ها

<sup>1</sup> Genetic Alogorithm

<sup>2</sup> Genetic Programing

۴- تعیین پارامترهای کنترل کننده اجرای برنامه (اندازه جمعیت، احتمال مربوط به به کارگیری عملهای ژنتیکی و جزئیات دیگر مربوط به اجرای برنامه)

۵- معیار پایان و ارائه نتایج اجرای برنامه (مثل تعداد تولید جمعیت جدید، تعیین یک مقدار مشخص برای برازش فرمول‌ها که اگر میزان برازش برابر یا بیشتر از آن مقدار شد، اجرا متوقف شود) در شکل ب-۱ طرح کلی گام‌های مقدماتی برنامه ریزی بیان ژن نشان داده شده است.



شکل ب-۱ طرح کلی گام‌های مقدماتی برنامه نویسی بیان ژن [۵۶]

اولین مرحله در الگوریتم GEP تولید جمعیت اولیه از راه حل هاست. این موضوع می‌تواند به وسیله فرآیند تصادفی و یا با استفاده از مقداری اطلاعات درباره مسئله انجام شود. سپس کروموزوم‌ها به صورت بیان درختی نشان داده شده، که این هم مطابق با یک تابع برازش ارزیابی می‌گردد. معمولاً تابع برازش به وسیله پردازش تعدادی از مسئله هدف، که مورد برازش نیز نامیده می‌شوند، ارزیابی می‌گردد. اگر کیفیت رضایت بخش از یک راه حل پیدا شود و یا نسل‌ها به تعداد معینی برسند، تکامل متوقف می‌شود و بهترین راه حل یافت شده تا به حال، گزارش داده می‌شود. از طرف دیگر اگر شرایط توقف یافت نشود، بهترین راه حل از نسل حاضر نگه داشته می‌شود (نخبه‌گزینی) و بقیه راه حل‌ها به فرآیند گزینشی واگذار می‌شوند. بر اساس گزینش، بهترین افراد شانس بیشتری برای تولید فرزندان

دارند. کل این روند برای چندین نسل تکرار می‌شود و با پیش رفتن نسل به جلو، انتظار می‌رود که کیفیت جمعیت نیز به طور متوسط بهبود یابد. پس بطور کلی می‌توان گفت که GEP از ژنوم خطی به عنوان پایه و اساس ژنتیکی و عملگرهای مانند جهش<sup>۱</sup>، تلفیق<sup>۲</sup>، نوترکیبی<sup>۳</sup>، وارونگی<sup>۴</sup> و جابجایی<sup>۵</sup> استفاده می‌کند. ژنوم متشکل از کروموزوم است و هر کروموزوم از ژن‌ها تشکیل شده است. که برای حل یک مساله به درخت بیان ترجمه شده است. فرآیند اجرایی برنامه ریزی بیان ژن به صورت مراحل زیر است [۵۷]:

۱- طول کروموزوم توسط تعدادی از ژن‌های آن تعریف می‌شود و طول ژن  $h + t$  تشکیل می‌شود.  $h$  پارامتر تعریف شده توسط کاربر است و  $t$  با این روش  $t = h(n-1) + 1$  حاصل می‌شود.  $n$  نشان‌دهنده ماکزیمم گره کاربردی در عبارت است.

۲- تولید یک جمعیت اولیه از فرمول‌ها که این فرمول‌ها از ترکیب تصادفی مجموعه توابع (عملگرهای ریاضی مورد استفاده در فرمول‌ها) و ترمینال‌ها (متغیرهای مسئله و اعداد ثابت) ایجاد می‌شوند، هر یک از افراد جمعیت مذکور با استفاده از توابع برازش مورد ارزیابی قرار می‌گیرند.

۳- عملگر جهش عباراتی را در طول ژنوم، جایگزین می‌کند. اگرچه باید به قواعد ژن مراجعه شود مانند توابع و نودهای ترمینال، آن‌ها در سرآیند<sup>۶</sup> ژن جهش می‌یابند، درحالی‌که نودهای ترمینال در انتها<sup>۷</sup> ژن جایگزین می‌شوند.

۴- تلفیق بین دو والد انتخاب شده از جمعیت رخ می‌دهد و بر اساس تلفیق تک نقطه، دو نقطه و تلفیق یکنواخت است. معمولاً نرخ جهش پایین است و با روش‌های کلاسیک مانند چرخ رولت و یا مسابقه‌ای انتخاب می‌شود و نرخ تلفیق معمولاً ۰/۷ است.

<sup>1</sup> Mutation

<sup>2</sup> Crossover

<sup>3</sup> Recombination

<sup>4</sup> Inversion

<sup>5</sup> Transposition

<sup>6</sup> Head

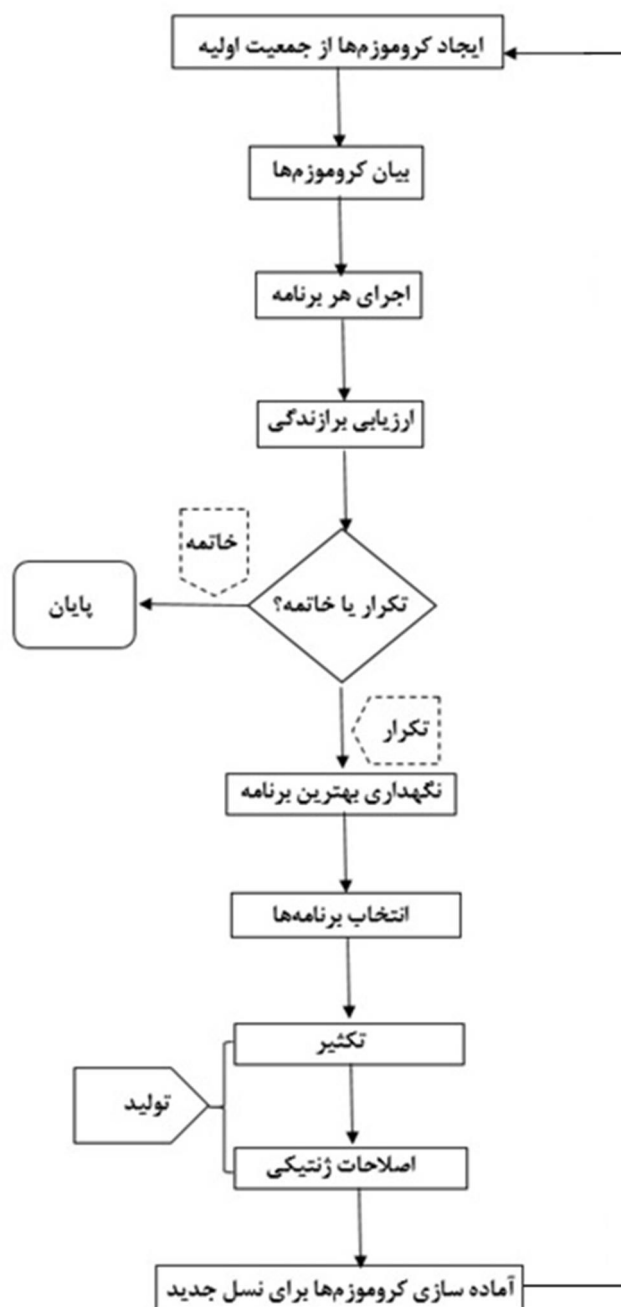
<sup>7</sup> Tail

۵- عملگر وارونگی ممکن است با احتمال کم در جهت معکوس توالی کوچکی از سمبل‌ها در بخشی از سرآیند یا انتهای ژن استفاده شود.

۶- عملگر جابجایی شامل حالت مختلفی است: کپی توالی قسمتی از ژن بروی سرآیند ژن، ریشه ژن، و یا انتقال کامل ژن در کروموزوم است. مراحل اصلی برنامه‌ریزی بیان ژن به طور شماتیک در شکل ب-۲ نشان داده شده است. یکی از موارد مهم در GEP تعیین تابع برازش است و هدف آن، یافتن راه حلی است که برای تمامی موارد برازش به اندازه یک خطای معین به خوبی عمل کند. از لحاظ ریاضی، برازش  $f_i$  از یک برنامه انفرادی  $i$  به صورت رابطه زیر بیان می‌شود.

$$f_i = \sum_{j=1}^n \left( R - \left| 100 \times \frac{P_{ij} - T_j}{T_j} \right| \right) \quad (1-1)$$

که در آن  $R$ ، محدوده انتخابی  $P_{ij}$  مقدار پیش‌بینی شده به وسیله برنامه انفرادی  $i$  برای مورد برازش  $j$  (از میان  $n$  مورد برازش) و  $T_j$  مقدار هدف برای مورد برازش  $j$  است. قابل توجه است که عبارت داخل قدر مطلق متناظر با درصد خطای نسبی است که این عبارت را دقت می‌نامند اگر  $T_j = P_{ij}$  باشد، خطا صفر می‌شود. بنابراین برای یک برازش کامل که در آن  $T_j = P_{ij}$  است.

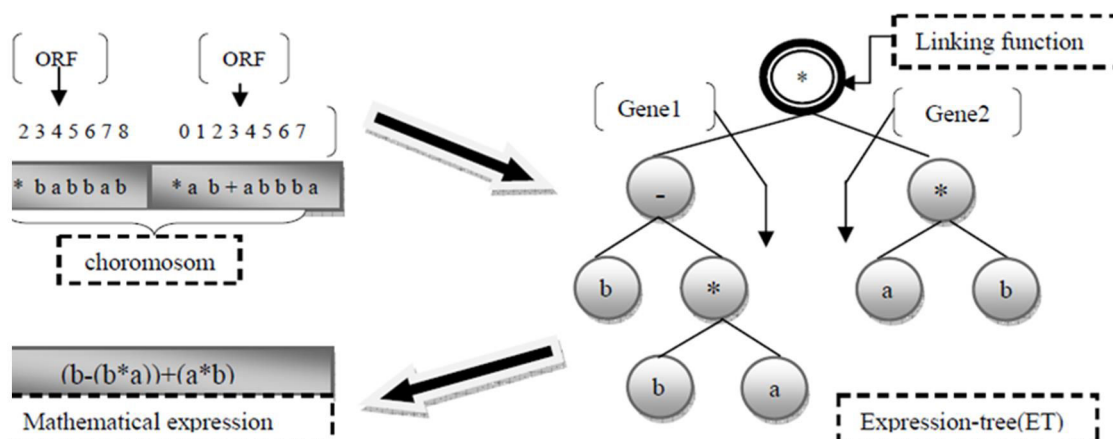


شکل ب-۲ فلوجارت برنامه‌ریزی بیان ژن [۵۷]

در GEP، هر ژن به صورت بیان درختی کدگذاری می‌شود. در مورد کروموزوم‌های چند ژنی، تمامی بیان درختی‌ها با استفاده از تابع پیوند، از محل گره ریشه خود به یکدیگر متصل می‌شوند. هر ژن، ناحیه کدگذاری به نام ORF (چارچوب تفسیر باز) دارد، که بعد از کد برداری (رمزگشایی) به صورت



ET بیان می‌شود و نشان‌دهنده یک راه‌حل کاندید، برای مسئله است شکل ب-۳ نشان می‌دهد چگونه یک کروموزوم با دو ژن به صورت یک رشته خطی کدگذاری شده و چطور به صورت ET بیان شده است. قابل ذکر است در این مثال هر دو ژن، درست مثل توالی‌های ژن‌های بیولوژیکی، نواحی کدگذاری شده و کدگذاری نشده دارند.



شکل ب-۳ کدگذاری به فرم رشته خطی و کد برداری به صورت ET کروموزوم با دو ژن در GEP [۵۵]

در GEP، کروموزوم‌ها طول ثابتی دارند و از یک یا چند ژن با طول مساوی تشکیل شده‌اند. بنابراین آنچه تغییر می‌کند طول ژن‌ها نیست، بلکه طول ORF است که تغییر می‌کند. در واقع طول ORF ممکن است مساوی یا کمتر از طول ژن باشد. اگرچه در GEP، محل شروع همیشه اولین نقطه از یک ژن است، ولی نقطه انتهایی همیشه منطبق با آخرین محل یک ژن نیست و نواحی کدگذاری نشده در قسمت بعد از نقطه انتهایی وجود دارد که در حقیقت، ماهیت GEP و قدرت استنتاج آن هستند. چون اجازه اصلاحات و به‌سازی ژنوم‌ها را بدون محدودیت، با استفاده از هر عملگر ژنتیکی می‌دهند و همیشه از نظر قواعد برنامه‌های صحیحی را تولید می‌کنند، بدون این‌که نیاز به فرآیند ویرایشی پیچیده یا روش‌های تحمیلی در جهت اجرای عملگرهای ژنتیکی داشته باشند. در حقیقت این بزرگ‌ترین تفاوت بین GEP و GP با یا بدون ژنوم‌های خطی است.



# پیوست ج

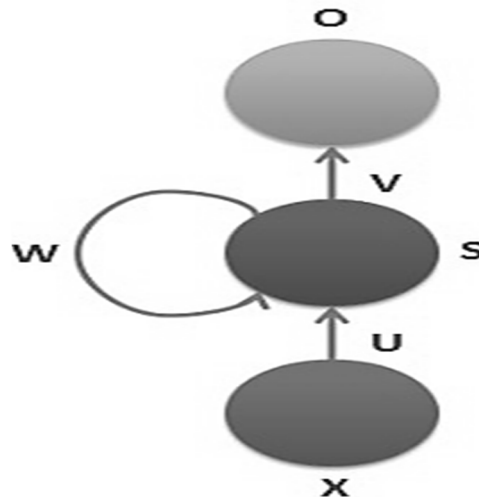
## یادگیری عمیق

حضور شبکه‌های LSTM در سال‌های اخیر بسیار پررنگ شده است. این شبکه‌ها در حال حاضر در فناوری‌های پیشرفته اعم از فناوری Google Voice استفاده می‌شوند اما تاریخچه آن به ۱۹۹۵ برمی‌گردد. LSTM (به معنی حافظه طولانی کوتاه-مدت) مخفف کلمه Long short-term memory یک نوع مدل یا ساختار برای داده‌های ترتیبی است که در سال ۱۹۹۵ برای توسعه شبکه‌های عصبی بازگشتی (RNN)<sup>۱</sup> ظهور پیدا کرد. در LSTM عبارت long term memory به وزن‌های یاد گرفته شده و short term memory به حالت‌های درونی سلول‌ها اطلاق می‌شود. LSTM برای حل مشکل پدیده ناپدید شده گرادیان در شبکه‌های عصبی بازگشتی به وجود آمدند که تغییر عمده آن جایگزین کردن لایه میانی RNN با یک بلاک (که بلاک LSTM نام دارد) است. اما قبل از اینکه به توضیح بیشتر در مورد شبکه LSTM بپردازیم ابتدا شبکه عصبی بازگشتی را توضیح داده سپس به توضیحات تکمیلی در مورد LSTM بپردازیم.

شبکه‌های عصبی متداولی که تاکنون متخصصان یادگیری ماشین از آن‌ها استفاده می‌کردند، نمی‌توانستند شبیه انسان عمل کنند و این یک نقصان بزرگ برای این شبکه‌ها محسوب می‌شد. بنابراین شبکه‌های عصبی بازگشتی برای برطرف کردن این مشکل طراحی شدند. در حقیقت شبکه‌های عصبی بازگشتی درون خودشان یک حلقه بازگشتی هستند که منجر می‌شود اطلاعاتی که از لحظات قبلی به دست آمده است از بین نرفته و در درون شبکه باقی بمانند. این شبکه‌ها در واقع برای پردازش سیگنال‌های دنباله‌دار به وجود آمدند. در یک شبکه عصبی معمولی تمام ورودی‌ها و خروجی‌ها مستقل از یکدیگر هستند، اما در بسیاری از موارد این ایده می‌تواند خیلی بد باشد. به عنوان مثال فرض کنید شما در یک جمله به دنبال پیش‌بینی کلمه بعدی هستید در صورتی که شبکه نتواند روابط بین کلمات را یاد بگیرد مسلماً نمی‌تواند کلمه بعدی را به درستی پیش‌بینی کند. اجازه دهید با یک دیدگاه دیگر به این نوع شبکه نگاه کنیم، این شبکه‌ها دارای یک نوع حافظه هستند که اطلاعاتی تاکنون دیده است را ضبط می‌کند. در تئوری این‌طور به نظر می‌رسد که این شبکه‌ها

<sup>۱</sup> Recurrent Neural Network

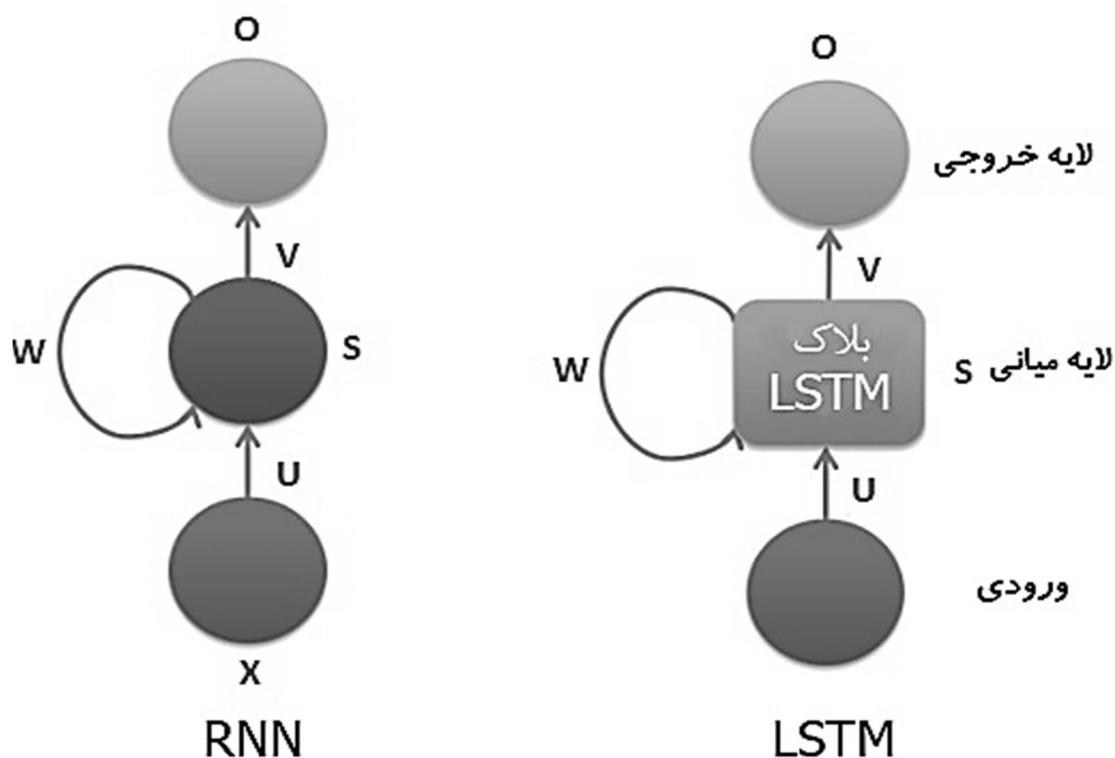
می‌توانند اطلاعات موجود در یک دنباله طولانی را ضبط و از آن‌ها استفاده کنند اما در عمل این‌طور نیست و بسیار محدود هستند، به این صورت که فقط اطلاعات چند گام قبل را ضبط می‌کنند. در شکل زیر نمونه‌ای از یک RNN معمولی نمایش داده شده است.



شکل ج-۱ ساختار یک شبکه عصبی بازگشتی [۵۸]

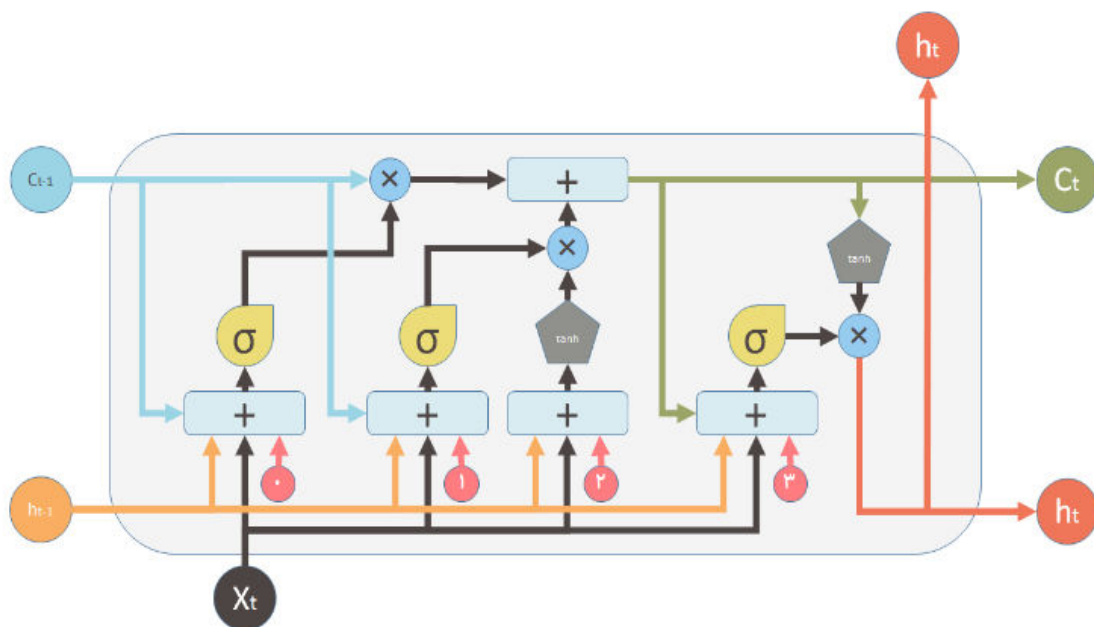
$X$  ورودی در گام زمانی  $t$  است، این ورودی می‌تواند یک بردار one-hot باشد (one-hot برداری است در یک بعد ۱ و در بقیه ابعاد صفر است).  $S$  حالت پنهان در گام زمانی  $t$  است. این قسمت همان جایی است که حافظه شبکه در آن قرار گرفته است.  $S$  بر اساس حالت پنهان قبلی و ورودی که در زمان جاری به آن داده می‌شود تغییر می‌کند که به شکل  $S(t) = f(Ux(t) + W(S(t-1)))$  است. تابع  $f$  معمولاً یک تابع غیرخطی مثل  $\tanh$  یا  $\text{ReLU}$  است.  $O$  خروجی در گام زمانی  $t$  است. به عنوان مثال در محاسبه کلمه بعدی در یک جمله این خروجی می‌تواند یک بردار از احتمالات در واژگان ما باشد. یکی از قابلیت‌های شبکه‌های عصبی بازگشتی این است که آن‌ها ممکن است بتوانند اطلاعاتی را که قبلاً مشاهده شده را به کاری که در حال حاضر در حال انجام است مرتبط سازد، برای مثال می‌توانند با استفاده از فریم‌های قبلی یک ویدئو، فریم کنونی را متوجه شوند. یا اینکه فرض کنید ما مدل زبانی‌ای ساخته‌ایم که تلاش می‌کند کلمه بعدی را با توجه به کلمات قبلی‌ای که در اختیارش قرار دادیم پیش‌بینی کند. اگر ما می‌خواهیم آخرین کلمه تو جمله «ابرها هستند در آسمان» رو پیش‌بینی کنیم، به اطلاعات اضافی دیگری نیاز نداریم و تقریباً می‌شود گفت واضح است که کلمه بعدی

«آسمان» است. در موارد مشابه این مثال، که فاصله بین اطلاعات مرتبط و جایی که به این اطلاعات نیاز داریم خیلی کم است، شبکه‌های عصبی بازگشتی می‌توانند یاد بگیرند که از این اطلاعات استفاده کنند. ولی ممکن است مواردی وجود داشته باشد که ما به اطلاعات بیشتری نیاز داشته باشیم. متأسفانه، هر چه این فاصله افزایش پیدا کند، شبکه‌های عصبی بازگشتی قدرتشان در به یاد آوردن و استفاده از اطلاعاتی که در گذشته دورتر یاد گرفته‌اند کاهش پیدا می‌کند و به عبارتی توانایی استفاده از اطلاعات گذشته دورتر را ندارند. شبکه‌های LSTM که نوع خاصی از شبکه‌های عصبی بازگشتی هستند که توانایی یادگیری وابستگی‌های بلندمدت را دارند. در حقیقت هدف از طراحی شبکه‌های LSTM، حل کردن مشکل وابستگی بلندمدت بود. به این نکته مهم توجه کنید که به یادسپاری اطلاعات برای بازه‌های زمانی بلند مدت، رفتار پیش فرض و عادی شبکه‌های LSTM است و ساختار آن‌ها به صورتی است که اطلاعات خیلی دور را به خوبی یاد می‌گیرند که این ویژگی در ساختار آن‌ها نهفته است. شکل ج-۲ ساختار یک شبکه LSTM را در مقایسه با شکل ج-۱ نشان می‌دهد.



شکل ج-۲ ساختار شبکه LSTM در برابر شبکه RNN [۵۸]

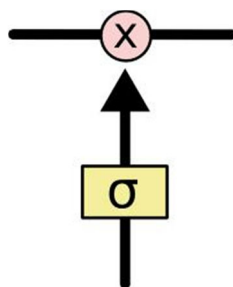
همه شبکه‌های عصبی بازگشتی به شکل دنباله‌ای (زنجیره‌ای) تکرار شونده از ماژول‌های (واحدهای) شبکه‌های عصبی هستند. در شبکه‌های عصبی بازگشتی استاندارد، این ماژول‌های تکرار شونده ساختار ساده‌ای دارند، برای مثال تنها شامل یک لایه تانژانت‌هایپربولیک ( $\tanh$ ) هستند. شبکه‌های LSTM نیز چنین ساختار دنباله یا زنجیره‌مانندی دارند ولی ماژول تکرار شونده ساختار متفاوتی دارد. به جای داشتن تنها یک لایه شبکه عصبی، ۴ لایه دارند که طبق ساختار ویژه‌ای با یکدیگر در تعامل و ارتباط هستند. شکل ج-۳ دیاگرام یک بلاک LSTM را نمایش می‌دهد. که در ادامه به صورت دقیق‌تر به تشریح هر یک از قسمت‌های آن خواهیم پرداخت.



شکل ج-۳ بلوک دیاگرام مربوط به LSTM [۵۸]

سلول حالت (cell state) عنصر اصلی LSTM است که در دیاگرام بالا یک خط افقی است که در بالای شکل قرار دارد (منظور خطی است که ابتدای آن  $C_{t-1}$  و انتهای آن را  $C_t$  مشخص کرده است). سلول حالت را می‌توان به صورت یک تسمه نقاله تصور کرد که از اول تا آخر دنباله یا همان زنجیره با تعاملات خطی جزئی در حرکت است یعنی ساختار آن بسیار ساده است و تغییرات کمی در آن اتفاق

می‌افتد. توسط ساختارهای دقیقی به نام دروازه‌ها<sup>۱</sup>، LSTM این توانایی را دارد که اطلاعات جدیدی را به سلول حالت اضافه یا اطلاعات آن را حذف کنید. این کار انجام می‌شود. دروازه‌ها راهی برای ورود اختیاری اطلاعات هستند. آن‌ها از یک لایه شبکه عصبی سیگموئید به همراه یک عملگر ضرب نقطه به نقطه تشکیل شده‌اند (شکل ج-۴).



شکل ج-۴ ساختار یک گیت [۵۸]

خروجی لایه سیگموئید عددی بین صفر و یک است، که نشان می‌دهد چه مقدار از ورودی باید به خروجی ارسال شود. مقدار صفر یعنی هیچ اطلاعاتی نباید به خروجی ارسال شود درحالی که مقدار یک یعنی تمام ورودی به خروجی ارسال شود. LSTM دارای ۳ دروازه مشابه برای کنترل مقدار سلول حالت است.

<sup>1</sup> Gates



## Abstract

The QA system is an automated system for obtaining the correct answers to questions posed by human in the natural language. In these systems, if the response is found, and if it is not the user's expected response or needs more information, there is no possibility of exchanging information between the system and the user to ask more questions and get answers related to it. To solve this problem, interactive Question answering (IQA) systems were created. Evaluation plays an important role in designing IQA systems. However, there is not yet a specific method for evaluating these systems in general, and it is only possible to take advantage of the evaluation methods used in the QA systems, dialog-based systems and the use of human assessors.

Therefore, providing a proper mechanism to the process of evaluating IQA systems can contribute significantly to improving these systems.

Presenting a model instead of human estimator can be a challenging task for assessing the interactive systems so that the output of the model can predict the score given by the estimator. In this thesis, an attempt was made to determine the appropriate parameters of a statistical model for evaluating IQA systems, and this statistical model can be used for assessing IQA systems in the assessment process. The aim of the statistical model was to provide the independence of the model from the language of the interactive system. In order to achieve the most appropriate model, several statistical features were extracted, then the regression and gene expression programming were used to reach the model.

First, a database of conversations took place with four IQA systems. Then, the feature extraction was performed on each conversation and finally the regression was used to extract the model. Also, due to the high number of

suggested features and to prevent over-fitting, the best features were chosen using REF method to form the proposed model based on the remaining features. The best model was determined by Lasso and Power series regression according to the root mean square error. Next, gene expression programming was used to achieve a more appropriate model. At the first step, a regression equation was formed to predict the score of the estimators based on the entire conversation. Then conversations were grouped into three classes with good, moderate and poor scores and a regression equation was obtained for each class. Therefore, for a new conversation, after the feature extraction phase and its allocation to a class, it was calculated based on the regression model.

According to the evaluation criteria, the average output error rate of the model compared to the actual output was 0.09, which indicated that the appropriate model was proposed in order to evaluate these systems. The proposed method in this study revealed that if a standard set of conversation between users and the system is available, it is the model's advantage to predict the score for the IQA systems if a human estimator is absent.

**Keywords:** Interactive Question Answering, Evaluation of Systems, Question Answering system, Regression, Gene expression programming.



**Kharazmi International Campus**

**PhD. Thesis in Artificial Intelligence Engineering**

**A Statistical Model for Evaluation of Interactive Question  
Answering Systems**

**By: Mohammad Mehdi Hosseini**

**Supervisor:**

**Dr. Morteza Zahedi**

**Advisor:**

**Dr. Hamid Hassanpour**

**July 2018**