

Presenting a Text Mining Algorithm to Identify Emotion in Persian Corpus

***Masoud Garshasbi¹, Anahid Rais-Rohani²,
Mohammadreza Kabaranzadeh Ghadim³***

Abstract: The literature regarding Persian text mining indicates that most studies are conducted to detect polarity of opinions on social websites. The aim of this research is presenting an algorithm to identify emotion implemented in the text based on the following six main emotions of happiness, sadness, fear, anger, surprise and disgust. In this research, the emotion will be examined based on unsupervised lexicon method. Identifying emotions conveyed by the texts based on a single emotional word will produce low accuracy because the intervening boosters and negating words can influence the emotion of the text too. Therefore, the algorithm has been implemented in six approaches with different features. In the first approach, the algorithm is capable of detecting only one emotional word in a sentence, and then it improves to detect boosters and negating and stop word list as well. The results of running the algorithm on two domains of data showed that the more features used in the algorithm, the more accurate the algorithm becomes and that the most effective factor is part of speech.

Key words: *Data mining, Emotion analysis, Sentiment mining, Text mining, Web mining.*

1. Research Instructor, Faculty of Iran Telecommunication Research Center, Tehran, Iran

2. MSc, Software Engineering, Islamic Azad University, Karaj Branch, Tehran, Iran

3. Associate Prof. of Management, Islamic Azad University, Central Tehran Branch, Tehran, Iran

Submitted: 09 / January / 2017

Accepted: 24 / September / 2017

Corresponding Author: Anahid Rais-Rohani

Email: anahidrr@gmail.com

ارائه الگوریتم متن کاوی به منظور تشخیص حس در متن های فارسی

مسعود گرشاسبی^۱، آناهید رئیس روحانی^۲، محمدرضا کابارن زاده قدیم^۳

چکیده: در متن کاوی متن های فارسی، در زمینه چگونگی استخراج ویژگی ها برای دسته بندی و بررسی نظرها در سایت های اجتماعی به منظور تشخیص قطبیت متن، مطالعاتی انجام شده است. هدف این پژوهش، ارائه الگوریتمی برای آنالیز حس متن فارسی، بر اساس شش حس پایه خوشحالی، ناراحتی، ترس، خشم، تعجب و تنفر است. در این پژوهش، آنالیز احساس به روش غیرنظارتی مبتنی بر لغتنامه انجام شده است. تشخیص حس جمله فقط با در نظر گرفتن یک لغت عاطفی دقت زیادی ندارد؛ زیرا عوامل دیگری نیز در جمله مانند تشدیدکننده ها و نفی کننده ها وجود دارند که روی حس متن تأثیر می گذارند. از این رو، الگوریتم به شش روش با در نظر گرفتن ویژگی های متفاوت نوشته شده است. در روش اول الگوریتم قابلیت تشخیص یک لغت عاطفی درون جمله را دارد؛ سپس قابلیت تشخیص تشدیدکننده، نفی کننده و لغات ایست اضافه می شود. نتایج به دست آمده از اجرای الگوریتم ها روی دو نمونه داده، نشان می دهد با در نظر گرفتن ویژگی های بیشتر، دقت الگوریتم نیز افزایش می یابد که در آن عامل قسمتی از سخن، بیشترین تأثیر را دارد.

واژه های کلیدی: آنالیز احساس، اندیشه کاوی، داده کاوی، متن کاوی، وب کاوی.

۱. مربی پژوهشی، پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران
۲. کارشناس ارشد مهندسی نرم افزار، دانشکده مکترونیک دانشگاه آزاد اسلامی واحد کرج، تهران، ایران
۳. دانشیار گروه مدیریت، دانشکده مدیریت، دانشگاه آزاد اسلامی واحد تهران مرکزی، تهران، ایران

تاریخ دریافت مقاله: ۱۳۹۵/۱۰/۲۰

تاریخ پذیرش نهایی مقاله: ۱۳۹۶/۰۷/۰۲

نویسنده مسئول مقاله: آناهید رئیس روحانی

E-mail: anahidrr@gmail.com

مقدمه

اندیشه کاوی یا آنالیز احساس زمینه مطالعاتی است که سعی می کند احساس ها، رفتارها، نظرها و تحلیل افراد مختلف را نسبت به موجودیت ها و ویژگی های آن بیان کند. این موجودیت می تواند محصول، سرویس، سازمان، فرد و رخداد و موضوع باشد. هدف از آنالیز یا تحلیل احساسات، پیدا کردن نظرهایی است که احساسی را نشان داده و قطبیت متن یا جهت گیری این نظرها را تشخیص می دهد.

افزایش اهمیت آنالیز احساسات با رشد رسانه های اجتماعی مانند نظرسنجی ها، فروم ها، انجمن های گفت و گو، وبلاگ ها، توئیتر و شبکه های اجتماعی همزمان شده است. سیستم های آنالیز احساسات کمابیش در همه زمینه های تجاری و اجتماعی به کار گرفته می شوند؛ زیرا نظرها و عقیده ها برای همه فعالیت های انسانی مهم بوده و تأثیر شایان توجهی بر رفتار ما دارند. یکی از مهم ترین کاربرد آنالیز احساس در علوم روان شناسی است. در سال های اخیر تحقیقات بسیاری برای پیش بینی خودکشی از طریق پست های درج شده در شبکه های اجتماعی انجام شده است.

حال با توجه به مطالعات انجام شده، چالش هایی در این زمینه برای زبان فارسی وجود دارد که از جمله آنها نقص پردازنده زبان فارسی، صرف شدن افعال، استفاده از وندهای تصریفی و حجم زیادی از کلمات محاوره ای است. برای آنالیز می توان از پردازنده های زبان انگلیسی نیز استفاده کرد، اما ابتدا باید این کلمات ترجمه شوند و خود ترجمه روی دقت آنالیز تأثیر منفی می گذارد.

در پژوهش حاضر، از آنجا که داده های استفاده شده ساختاریافته نیستند، الگوریتم پیشنهاد شده مبتنی بر روش لغتنامه پیاده سازی می شود و به منظور افزایش دقت این روش، علاوه بر لغات عاطفی، از ویژگی های دیگری نظیر تشدیدکننده و نفی کننده نیز بهره برده شده است تا این الگوریتم بتواند حس متن را بر اساس شش پایه تشخیص دهد.

نوآوری این مقاله استفاده نکردن از الگوریتم های زبان دیگر است؛ بدین ترتیب متن با ترجمه دچار تغییر حس نمی شود. همچنین برای آنالیز، از تعداد تکرار لغت های مربوط به دسته حس (منظور از دسته حس، شش حس پایه خوشحالی، ناراحتی، خشم، ترس، نفرت و تعجب است) که به آن تعلق دارند، استفاده نشده است؛ بلکه در لغتنامه به لغت ها وزنی اختصاص داده شده و با محاسبه نهایی آن و تأثیر مؤلفه های دیگر در جمله، حس آن مشخص شده است.

در ادامه، ابتدا پژوهش های گذشته مرور می شود؛ سپس چگونگی تهیه لغتنامه ها و شرحی از مراحل اجرای الگوریتم پیشنهاد شده توضیح داده می شود و در آخر ضمن بحث و نتیجه گیری درباره اجرای الگوریتم روی متن های فارسی، مقایسه ای با الگوریتم نایوبیز صورت می گیرد.

پیشینه پژوهش

آنالیز احساس که شاخه‌ای از متن‌کاوی محسوب می‌شود، در حال حاضر به موضوع روز تبدیل شده و تحقیقات آن در زبان‌های خارجی از سال ۲۰۰۰ آغاز شده است. در زبان فارسی، تحقیقات بر چگونگی استخراج ویژگی و تغییر ویژگی‌ها برای افزایش دقت الگوریتم‌ها متمرکز است. در ادامه، توضیح کوتاهی از کارهای انجام شده با روش‌ها و دامنه‌های مختلف در این زمینه ارائه می‌شود.

یو، وو و یانگ (۲۰۰۹) روی اسناد سایت مشاوره دو مدل بازیابی مبتنی بر لغت (مدل و مدل OKAPI^۲) را با مدل مبتنی بر گفتمان مطلع مقایسه کردند. نتیجه نشان داد مدل بازیابی اطلاعات گفتمان نسبت به مدل بازیابی مبتنی بر لغت، دقت بیشتری دارد.

هادانوا، کاشیادا و اندوا (۲۰۱۱) به صورت دستی، جنبه‌ای از احساسات را در اسناد بازیابی یکی از بازی‌های استخراج شده از وبسایت شناسایی کردند. آنها در این تحقیق از روش خوشه‌بندی جملات و ابزار Bayon مبتنی بر الگوریتم دوبرخی استفاده کردند و به این نتیجه رسیدند که روش خوشه‌بندی بهتر از روش بدون خوشه‌بندی است.

لامبوا، پایشا و دیاسا (۲۰۱۱)، چهار نوع الگوریتم توافقی شامل SAR^۳، MAA^۴، BMAA^۵ و BMAADR^۶ را ادغام کردند و نتیجه گرفتند که الگوریتم SAR بهترین نتیجه را تولید کرده است. روی مجموعه تفسیرهای دستی RIMDB، میانگین دقت الگوریتم SAR حدود ۷۷/۱ درصد بود. بهترین اجرا الگوریتم MAA دقتی به اندازه ۷۵/۶ داشت، اما این میزان دقت از بهترین اجرای الگوریتم SAR بدتر بود. میانگین دقت الگوریتم BMAADR نیز حدود ۸۰ درصد به دست آمد. به طور کلی، الگوریتم SAR اجرای بهتری را در حالت‌های مجموعه داده منحصرأ ذهنی و عینی دارد و الگوریتم BMAADR بهترین اجرای الگوریتم برای متن‌های واقعی دارد است.

هادیا، هولیو و شیب (۲۰۱۳) به بررسی تأثیر پیش‌پردازش روی متن‌های گرفته شده از اینترنت پرداختند که نتیجه آن به افزایش دقت ماشین بردار منجر شد. برای داده‌های

-
1. Viable System Model
 2. Okapi is the name of an animal related to zebra, the system where this model was first implemented was called Okapi
 3. Synthetic Aperture Radar
 4. Message Authenticator Algorithm
 5. L-Beta-methylamino-alanine
 6. Balanced Merged Agreement Algorithm Using Documents Rank

پیش پردازش نشده، دقت در ماتریس تکرار ویژگی^۱ پیشرفت داشت و برای داده های پیش پردازش شده دقت هم در تکرار ویژگی و هم در حضور ویژگی^۲ افزایش یافت.

دوتی (۲۰۱۳) روی متن های ساده خبری انگلیسی، توابعی را برای تشخیص حس ارائه داد که اغلب مربوط به تشخیص نادرست توکن در برنامه برای false positive بود. وی از ماژول رفع ابهام و برجسبزن قسمتی از سخن^۳ استفاده کرد و موجب پیشرفت فراخوانی کلی شد.

باراوی هاردیمن و سنگ (۲۰۱۳) نیز روی ارزیابی دقت منابع، آزمایش هایی انجام دادند. روی ویدئوهای آموزشی و با اندازه فاصله اقلیدسی برای مقایسه نتایج به دست آمده از آنالیز توصیفی استفاده کردند و متوجه شدند که با رسم نمودار تولید، منابع مناسب برای تشخیص احساسات در نظرهای مربوط به ویدئوهای یادگیری آنلاین را می توان تفسیر کرد.

بالاهور و تورچی (۲۰۱۴) به آزمایش مقایسه ای با استفاده از یادگیری تحت نظارت و ماشین ترجمه به منظور تجزیه و تحلیل احساسات چندزبانه اقدام کردند. برای ترجمه از چهار سیستم SMT^۴، گوگل، بینگ، mooses و برای دسته بندی از ماشین بردار دسته ای^۵ استفاده کردند و نشان دادند استفاده از ماشین بردار دسته ای، تأثیر مثبتی بر نتایج می گذارد.

قاضی، اینکین و اسپاکویچ (۲۰۱۴) با استفاده از روش لغتنامه به دسته بندی متن به شش حس پایه پرداختند. در این پژوهش سه مدل پایه، ماشین بردار و بهینه سازی حداقل پی در پی^۶ در نرم افزار وکا بررسی شده است. بر اساس نتایج به دست آمده، ماشین بردار روی دیتاست بزرگ نتیجه بهتری داشت و LR^۷ در مقایسه با ماشین بردار روی ویژگی های مطرح شده در دسته ای از لغات بهتر جواب داد.

بانا، میالسی و ویپ (۲۰۱۴) دو روش چندزبانه و زبان ضربداری^۸ را مقایسه کردند و نتیجه گرفتند که دقت روش چند زبانه از زبان ضربداری بیشتر است.

بریخن، کونوپیک (۲۰۱۴) الگوریتم های هال^۹، کالز^{۱۰}، بیگل^{۱۱} و اصول و پارامترها^{۱۲} را روی متن های اخباری به زبان های انگلیسی، اسلو و چک آزمایش کردند. بر اساس نتایج آنها، روش

-
1. Feature Frequency (FF)
 2. Feature Presence (FP)
 3. Part Of Speech (POS)
 4. Satisfiability Modulo Theories
 5. Bagging SVM
 6. Sequential Minimal Optimization (SMO)
 7. Left to Right approach
 8. Crosslingual
 9. Hal
 10. Coals
 11. Beagle
 12. Principles and Parameters (P&P)

حال برای خوشه‌های متراکم و روش کالز برای خوشه‌های کم‌پشت مناسب است، اما به‌طور کلی ترکیب حال و کالز بهترین نتیجه را می‌دهد. همچنین، الگوریتم‌های بیگل و اصول و پارامترها در مقابل روش پایه پیشرفت قابل اغمازی دارند.

علیمردانی و آقای (۲۰۱۵) روش الگوریتم ماشین بردار را با لغتنامه ترکیب کردند و آن را روی نظرهای ارائه شده در خصوص هتل با چهار فرضیه تعداد تکرار کلمه‌ها، حضورداشتن و حضورنداشت کلمه‌ها، حاصل ضرب تعداد تکرار در قطبیت و حاصل ضرب حضور داشتن و حضور نداشتن کلمه‌ها در قطبیت، به اجرا درآوردند. نتایج نشان داد فرضیه حاصل ضرب تعداد تکرار در قطبیت، بهترین نتیجه را دارد.

دویکا، سونیتا و گانشا (۲۰۱۶) روش‌های مختلف آنالیز احساس را بررسی کردند. نتایج این مقایسه نشان داد در روش ماشین یادگیری به لغتنامه نیازی نیست، اما یکی از معایب آن وابستگی به دامنه است. روش مبتنی بر قانون دقت زیادی دارد، اما این دقت به تعریف قوانین وابسته است. روش مبتنی بر لغتنامه به داده‌های برجسب خورده احتیاجی ندارد، بلکه به لغتنامه غنی نیاز دارد که ممکن است همیشه در دسترس نباشد.

روش‌شناسی پژوهش

در روش غیرنظارتی از لغتنامه و فرهنگ لغت استفاده می‌شود و مجموعه قوانینی برای محاسبه نتیجه وجود دارد. برای این کار به لغتنامه فارسی نیاز است و چون در این زمینه لغتنامه فارسی وجود نداشت، لغتنامه‌هایی در قالب فایل اکسل تهیه شد که در ادامه به شرح آن پرداخته می‌شود.

لغتنامه‌ها

لغتنامه احساس: این لغتنامه شامل فهرست لغت‌هایی است که به یکی از شش حس پایه بیان شده در پژوهش قاضی و همکارانش (۲۰۱۴) برجسب زده شده‌اند که عبارت‌اند از: شاد، غمگین، تعجب، خشم، تنفر و ترس. در ادامه، به هر یک از این لغت‌ها وزن عددی بین ۱ تا ۴ مبنی بر مقدار اثرگذاری برای تشخیص صحیح حس، اختصاص داده شده است. جدول ۱ نمونه‌ای از این لغتنامه را نشان می‌دهد.

لغتنامه تشدیدکننده: این لغتنامه شامل فهرست لغت‌های تشدیدکننده با برجسب قطبیت منفی یا مثبت است و قطبیت آنها دارای وزن عددی بین ۱ تا ۴ است که میزان تأثیر لغت‌ها را نشان می‌دهد. جدول ۲ نمونه‌ای از این لغتنامه را به نمایش گذاشته است.

جدول ۱. نمونه لغتنامه احساس

حس	وزن	لغت
خوشحالی	۴	موفق
ناراحتی	۴	مشکل
عصبانیت	۲	لااقل
خوشحالی	۳	سپاس
ترس	۳	مبادا
تعجب	۳	جدی

جدول ۲. نمونه لغتنامه تشدیدکننده

وزن	لغت
۳	بسیار
۳	خیلی
-۴	هیچ
-۲	کمی
۴	کاملاً

لغتنامه نفی: این لغتنامه فهرستی از لغتهایی را دربرمی گیرد که نفی کننده اند و فعل یا پیشوندی با بار منفی دارند. در این لغتنامه هر لغت با برچسب v و n مشخص شده است. v به معنای فعل نفی یا پیشوند فعل نفی بوده و n پیشوند نفی اسم یا صفت است. جدول ۳ نمونه ای از لغتنامه نفی را نشان می دهد.

جدول ۳. نمونه لغتنامه نفی

برچسب	لغت
v	نمی
n	نه
n	عدم
n	بی
v	نیست
v	نبود
v	نکن
v	ندانم

لغتنامه توقف: در این لغتنامه فهرستی از لغت‌ها و حروف اضافه قرار دارد که هیچ‌گونه حسی را بیان نمی‌کنند و تأثیری بر حس متن ندارند. لغات آن شامل روزهای هفته، اعداد، ضمائر، حروف اضافه و اسامی اشخاص و اشیا است.

جدول ۴. نمونه لغتنامه توقف

از
کتاب
برای
یا
شنبه
حمید
دویست
را
تو

داده‌ها

در آنالیز احساس به غنی بودن داده‌های انتخاب‌شده برای طبقه‌بندی باید توجه کرد. در پژوهش حاضر، داده‌ها از وبسایت اجتماعی آموزش الکترونیکی انتخاب شده است^۱. در این سامانه بخشی به نام آشار وجود دارد که در آن دانشجویان می‌توانند در قالب متن با کارشناسان پشتیبانی، استادان و دانشجویان هم‌گروه خود گفت‌وگو کنند.

با آنالیز احساس این مکاتبات، می‌توان میزان رضایتمندی دانشجویان از استادان، محتوای آموزشی و توقع دانشجو از سیستم مدیریت یادگیری الکترونیکی را ارزیابی کرد و به منظور ارتقای کیفیت آموزش الکترونیکی، افزایش خدمات مورد نیاز دانشجویان به صورت الکترونیکی و جلب رضایت بیشتر دانشجویان، برنامه‌های مدیریتی و توسعه‌ای تعیین و اولویت‌بندی کرد.

از آنجا که این داده‌ها مربوط به مکاتبات افراد بوده و ممکن است در محیط گفت‌وگو به زبان انگلیسی نوشته شده باشند؛ به صورت دستی جمله‌های مکتوب شده به زبان فارسی جدا شدند. از لغتنامه توقف نیز برای پاکسازی جمله‌ها از حروف اضافه و کلمات بی‌تأثیر استفاده شد.

ویژگی ها

انتخاب ویژگی در متن کاوی مرحله بسیار مهمی است. در این تحقیق از چهار دسته ویژگی، لغات عاطفی، قسمتی از سخن که دارای لغات تشدیدکننده و افعال نفی است، وابستگی بین لغات و همچنین تعداد لغات عاطفی در جمله استفاده شده است.

- لغات عاطفی: ویژگی ها بر خود لغات تکیه دارند. این دسته برای جمله های ساده مناسب است. این لغات در لغتنامه احساس وجود دارند و به هر یک برچسب حس زده شده است.
- قسمتی از سخن (ارکان جمله): لغاتی مانند تشدیدکننده و نفی کننده و همچنین افعال نفی هستند که در تشخیص حس جمله تأثیرگذارند. برای تشخیص آنها از لغتنامه تشدیدکننده و نفی کننده استفاده شده است.
- وابستگی لغات: در جمله هایی که کلمه تشدیدکننده روی حس لغت عاطفی تأثیر می گذارد یا پیشوندهای نفی کننده که روی لغات بعد از آن در جمله اثر دارد، در این دسته ویژگی ها قرار می گیرند. این ویژگی برای جمله هایی که بیش از یک لغت عاطفی دارد به کار می رود؛ زیرا ترتیب قرارگیری تشدیدکننده و نفی کننده مشخص می شود.
- تعداد لغات عاطفی در جمله: در جایی کاربرد دارد که جمله دربردارنده چند لغت عاطفی با برچسب حس متفاوت است. در این ویژگی، بیشترین مجموع وزن لغات در یک دسته مشابه، حس جمله را تعیین می کند.

یافته های پژوهش

پس از تهیه لغتنامه های بیان شده در بخش قبل، الگوریتم به زبان سی شارپ با شش روش ترکیبی استفاده از ویژگی ها پیاده سازی شد و برای بررسی عملکرد آنها ویرایشگری تهیه گردید تا جمله را به صورت ورودی از کاربر گرفته و با انتخاب هر یک از ویژگی ها تشخیص حس توسط الگوریتم را ارزیابی کند.

روش نخست: استفاده از ویژگی لغت

الگوریتم تنها یک لغت عاطفی در جمله را با استفاده از لغتنامه عاطفی تشخیص می دهد و حس آن لغت، حس جمله را بیان می کند. این روش برای جمله های ساده مناسب است. دقت این روش ۳۰ درصد محاسبه شد.

روش دوم: استفاده از ویژگی لغت عاطفی و تعداد لغات در جمله

در این روش علاوه بر ویژگی لغات عاطفی، از ویژگی تعداد لغات عاطفی در جمله نیز بهره برده شده است. در این روش نیز لغات با استفاده از لغتنامه عاطفی تشخیص داده می‌شوند، ولی این بار الگوریتم قابلیت تشخیص بیش از یک لغت عاطفی را دارد و حس آن لغتی که بیشترین وزن را دارد، به‌عنوان حس جمله شناسایی می‌کند. دقت الگوریتم در این روش ۴۰ درصد بود.

روش سوم: استفاده از ویژگی لغت عاطفی و قسمتی از سخن (ارکان جمله)

علاوه بر لغات عاطفی، لغات تشدیدکننده نیز به‌عنوان ویژگی در نظر گرفته می‌شوند. الگوریتم تنها یک لغت عاطفی و یک لغت تشدیدکننده در جمله را تشخیص می‌دهد. درصد وزن لغت تشدیدکننده به وزن لغت عاطفی اضافه شده، وزن نهایی را تعیین می‌کند. دقت این روش نیز مانند روش اول ۳۰ درصد محاسبه شد.

روش چهارم: استفاده از ویژگی لغت عاطفی، تعداد لغات در جمله و وابستگی لغات

الگوریتم بیش از یک لغت عاطفی و لغت تشدیدکننده درون جمله را تشخیص می‌دهد و ترتیب قرارگیری تشدیدکننده و لغات عاطفی را در نظر می‌گیرد و مشخص می‌کند تشدیدکننده برای کدام لغت عاطفی استفاده شده است. در این روش الگوریتم ابتدا وزن تشدیدکننده‌ای را که تشخیص داده، ذخیره می‌کند و پس از تشخیص نخستین لغت عاطفی، درصد آن را به وزن لغت اضافه کرده و مقدار نهایی آن دسته را محاسبه می‌کند و پس از محاسبه وزن تمام لغات، حس آن دسته که بیشترین وزن را به‌دست آورده، به‌عنوان حس جمله انتخاب می‌کند. الگوریتم در این روش دقت بالاتری داشت و ۵۰ درصد محاسبه شد.

روش پنجم: استفاده از ویژگی لغت عاطفی، تعداد لغات در جمله و قسمتی از سخن (ارکان جمله)

در این روش، فعل نفی نیز علاوه بر لغات عاطفی و تعداد لغات عاطفی به‌عنوان ویژگی در نظر گرفته می‌شود. در روش پنجم نیز مانند روش چهارم، ترتیب قرار گرفتن لغت عاطفی با تشدیدکننده اهمیت دارد و باید مشخص شود تشدیدکننده برای کدام لغت عاطفی استفاده شده تا بتوان وزن و قطبیت تشدیدکننده را روی آن لغت عاطفی خاص تأثیر داد. پس متغیری اضافه می‌شود تا لغت تشدیدکننده را با مقدار وزنی که دارد، ذخیره کند؛ سپس وزن آن روی لغت عاطفی بعدی‌ای که پیدا شده، تأثیر داده می‌شود. نتیجه این مرحله با وزن ماکزیمم جمله مقایسه

می شود، چنانچه مقدار بیشتری داشت، در وزن ماکزیمم جایگزین شده و حس آن به عنوان دسته حس جمله در نظر گرفته می شود.

این کار تا پایان جمله ادامه می یابد و اگر از فعل نفی کننده استفاده شده باشد، حس کل جمله را تحت تأثیر قرار می دهد؛ بنابراین، الگوریتم حس به دست آمده را معکوس می کند. برای مثال، اگر حس جمله جزء دسته خوشحالی بود، نفی آن جز دسته ناراحتی می شود. برای دسته حس هایی که معکوس وجود ندارد تا بتوان نفی آن حس را به یکی از دسته ها نسبت داد، یک دسته دیگری به عنوان خنثی در نظر گرفته شده است که حس لغت در این دسته قرار داده می شود. دقت الگوریتم در این روش به ۷۰ درصد رسید.

روش ششم: استفاده از ویژگی لغت عاطفی، تعداد لغات در جمله، قسمتی از سخن (ارکان جمله) و وابستگی

در این روش هر چهار ویژگی باهم استفاده می شود. مانند روش چهارم نه تنها ترتیب قرار گرفتن لغت عاطفی با تشدیدکننده اهمیت دارد، بلکه ترتیب نفی کننده ها نیز مهم است. در لغتنامه نفی کننده ها، برچسب زده شده روی هر یک مشخص می کند که لغت نفی کننده پیشوند است؛ برای مثال «بدون» یا فعل «نیست» که نفی کننده است.

جدول ۵. نمونه ای از جملات و نتایج به دست آمده در شش روش

جملات	روش ۱	روش ۲	روش ۳	روش ۴	روش ۵	روش ۶
با تشکر از شما استاد عزیز	خوشحالی	خوشحالی	خوشحالی	خوشحالی	خوشحالی	خوشحالی
خیلی ممنونم از شما ولی مشکلم حل نشد	خوشحالی	ناراحتی	خوشحالی	خوشحالی	ناراحتی	ناراحتی
ممنونم از شما کمی مشکلم حل شد	خوشحالی	ناراحتی	خوشحالی	خوشحالی	خوشحالی	خوشحالی
مشکلتان را دقیق تر بنویسید تست شد مشکلی نیست.	ناراحتی	ناراحتی	ناراحتی	ناراحتی	خوشحالی	خوشحالی
خیلی ممنونم از راهنماییتون متأسفانه پشتیبان متوجه منظور من نشده	خوشحالی	ناراحتی	خوشحالی	خوشحالی	ناراحتی	ناراحتی

برای این کار متغیری گذاشته می شود تا لغت نفی کننده را همراه با برچسب آن ذخیره کند؛ سپس مانند حالت قبل، اگر لغت بعدی لغت عاطفی باشد و برچسب نفی کننده پیشوند باشد، آن را روی حس لغت عاطفی تأثیر می دهد و اگر تشدیدکننده نیز در جمله وجود داشته باشد، پس از

یافتن تشدیدکننده، درصد وزن آن را به وزن لغت می‌افزاید؛ سپس وزن به‌دست آمده را با وزن ماکزیمم جمله مقایسه می‌کند، چنانچه مقدار بیشتری داشت، وزن حاصل را در وزن ماکزیمم و دسته آن را جزء دسته جمله در نظر می‌گیرد. این کار تا پایان جمله ادامه می‌یابد. چنانچه نفی‌کننده فعل باشد، پس از مشخص شدن وزن ماکزیمم و حس وزن ماکزیمم، آن فعل نفی را روی حس کل جمله تأثیر می‌دهد که تأثیر آن نیز مانند روش پنجم است. در این روش دقت الگوریتم پیشرفت کرد و به ۸۰ درصد رسید. خلاصه‌ای از آزمایش‌های انجام شده در جدول ۵ آورده شده است.

روش الگوریتم یادگیری نایو بیز

در این روش قسمتی از پیاده‌سازی در نرم‌افزار هوش مصنوعی مایکروسافت انجام گرفت و از برنامه SQL Business Inteligence استفاده شد. دیتاسورس یا همان داده‌های آموزشی، شامل ۱۰۰ نمونه داده‌های وب‌سایت اجتماعی آموزش الکترونیکی است که کاربران وارد کرده‌اند. روی این داده‌ها توسط الگوریتم روش ششم برچسب حس زدیم و نتیجه آن توسط مفسر بررسی و تصحیح شد؛ سپس آن را روی داده‌های وب‌سایت اجتماعی آموزش دادیم. بقیه پیاده‌سازی در برنامه نوشته شده به زبان سی شارپ، انجام گرفت. بدین ترتیب مدلی با تکنیکی که از آن تجربه کسب می‌کند، ساخته می‌شود. در این تحقیق، مدل بر اساس الگوریتم یادگیری نایو بیز ساخته شده است. ویژگی‌ها شامل لغات عاطفی، لغات تشدیدکننده و لغات نفی‌کننده مانند الگوریتم روش ششم مشخص می‌شوند؛ اما در این روش ویژگی‌ها به‌صورت مستقل هستند. برای این کار ستون‌هایی از جدول دیتاسورس به‌عنوان ورودی و یک ستون برای درج پیش‌بینی، در نظر گرفته شده است. در این برنامه کدی نوشته شد که به کمک آن می‌توان به برنامه نرم‌افزار هوش مصنوعی مایکروسافت متصل شد. سپس فرمی تهیه شده تا از طریق آن، الگوریتم را روی داده‌های جدید وب‌سایت اجرا کند و حس جمله را تخمین بزند.

نتیجه‌گیری و پیشنهادها

ابتدا نتایج به‌دست آمده از اجرای الگوریتم به شش روش مرور شده و در ادامه، الگوریتم روش ششم با اجرا روی دو دامنه متفاوت و همچنین با مقایسه الگوریتم نایو بیز ارزیابی می‌شود.

نتایج به‌دست آمده از شش روش

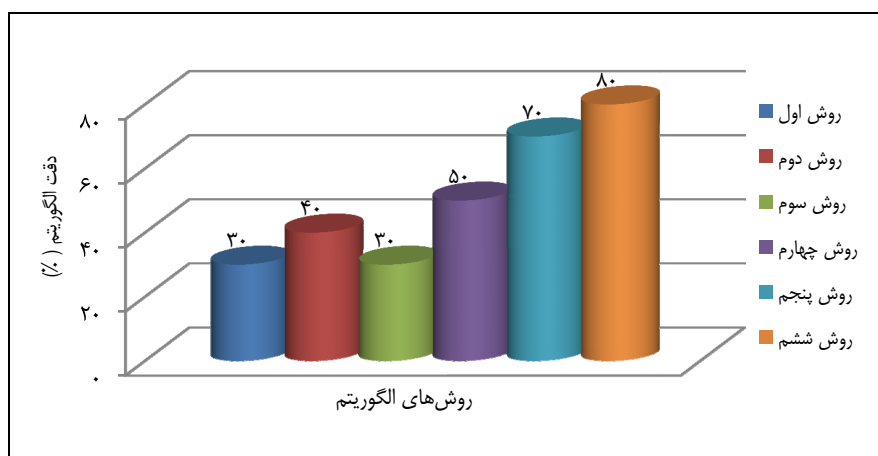
با گرفتن ایده از پژوهش هادیا و همکارانش (۲۰۱۳) که نشان دادند پیش‌پردازش تأثیر مثبتی بر دقت الگوریتم دارد و همچنین با بهره بردن از ویژگی‌هایی که در اجرای بهتر الگوریتم مقاله

قاضی و همکارانش (۲۰۱۴) مؤثرتر بودند، الگوریتم را به شش روش ارزیابی کردیم و در هر روش تعداد ویژگی ها را تغییر دادیم. حال به مقایسه نتایج پرداخته می شود.

برای مقایسه روش اول و دوم، جمله «دیروز شاگردم را تنبیه کردم و الان پشیمان و ناراحت هستم» بررسی شد. در روش اول که تنها لغات عاطفی در نظر گرفته می شود، الگوریتم با استخراج لغت تنبیه حس جمله را خشم تشخیص داد؛ اما در روش دوم که تعداد لغات عاطفی نیز در نظر گرفته می شود، تشخیص حس جمله، ناراحتی بود. دقت روش اول ۳۰ درصد و روش دوم ۴۰ درصد محاسبه شد که از تغییر درصد دقت می توان دریافت که ویژگی تعداد لغات عاطفی در جمله بر دقت آنالیز حس توسط الگوریتم، تأثیر مثبتی می گذارد.

برای مقایسه روش چهارم و پنجم، جمله «خیلی ممنونم اما مشکلم حل نشد» برای آزمایش انتخاب شد. در روش چهارم که ویژگی «وابستگی» مطرح است و از ویژگی «قسمتی از جمله» که ویژگی تشخیص فعل جمله است بهره نمی برد، حس جمله خوشحالی تشخیص داده شد و در روش پنجم که الگوریتم توانست فعل را شناسایی کرده و تأثیر آن را در جمله محاسبه کند، حس جمله ناراحتی بود. دقت روش چهارم تا ۵۰ درصد پیشرفت داشت که نسبت به روش های قبلی نشان می دهد ویژگی تعداد لغات عاطفی تأثیر بیشتری دارد؛ اما در مقایسه روش چهارم و روش پنجم که دقت آن ۷۰ درصد بود، می توان گفت ویژگی قسمتی از سخن، به مراتب تأثیر گذارتر از ویژگی وابستگی است.

نتیجه دقت به دست آمده از اجرای الگوریتم به روش ششم که در آن الگوریتم از چهار ویژگی بهره برده است، به ۸۰ درصد رسید.



شکل ۱. مقایسه روش های انجام شده روی داده های کاربران

ارزیابی الگوریتم

برای ارزیابی الگوریتم به روش ششم، از دو روش زیر استفاده شد. از الگوریتم روش ششم، برای تشخیص حس در داده‌های مکاتبه‌ای افراد استفاده شد. نتایج به‌دست آمده از ۵۰ نمونه داده شده به الگوریتم پیشنهادی به این صورت بود: ۸۴ درصد حس جمله تشخیص درستی داشت؛ ۴ درصد جملات بدون حس را به اشتباه دارای حس شناسایی کرد و ۱۲ درصد جملات دارای احساس را نادرست تشخیص داد. مقایسه‌ای نیز بین دو روش با داده‌های کاربران و داده‌های وبسایت اجتماعی صورت گرفت. نتایج نشان داد دقت الگوریتم با داده‌های کاربران ۸۰ درصد و با داده‌های وبسایت اجتماعی ۸۴ درصد است.

دقت الگوریتم با داده‌های وبسایت بیشتر بود؛ زیرا لغتنامه‌ها از لغات استفاده شده در وبسایت تهیه شده‌اند. در نتیجه می‌توان گفت که در روش لغتنامه‌ای، دامنه داده در دقت تشخیص حس توسط الگوریتم تأثیر می‌گذارد؛ اما ممکن است در جملات از صفاتی استفاده شود که در لغتنامه وجود ندارد که این یکی از مشکلات روش لغتنامه است. پس هرچه لغتنامه غنی‌تر باشد خطا کمتر است.

همچنین الگوریتم پیشنهاد شده با الگوریتم یادگیری نایویز مقایسه شد. این بار الگوریتم‌ها روی ۵۰ نمونه داده دانشجویان در وبسایت اجتماعی آموزش الکترونیکی به‌طور یکسان اجرا شدند. در الگوریتم پیشنهاد شده، ۷۸ درصد حس جمله تشخیص درست و ۲۲ درصد جملات تشخیص اشتباه داشت. در روش یادگیری که از الگوریتم نایویز استفاده شده است، ۵۲ درصد حس جمله تشخیص درست و ۴۸ درصد جملات تشخیص نادرست داشت. برای بهبود روش یادگیری با الگوریتم نایویز، پیشنهاد می‌شود حجم داده‌های یادگیری بیشتر شود.

در این تحقیق ادعا شد که معنای اولیه لغات برای تشخیص حس جمله کافی نیست و سایر لغات در جمله نیز روی حس جمله تأثیرگذارند. برای ارتقای الگوریتم، به تشخیص قیود و نفی‌کننده‌ها نیز پرداخته شد. در این تجربه‌ها، الگوریتم طراحی شده با ترکیب مختلف ویژگی‌ها ارزیابی شد و در نتیجه آن مشخص گردید که ترکیب چهار ویژگی تأثیر بیشتری برای تشخیص درست حس جمله دارد.

همچنین برای رفع چالش صرف افعال، افعال به‌صورت صرف شده در لغتنامه درج شدند. شایان ذکر است که پیشوندهای نفی نیز در لغتنامه نفی اضافه شده بود و با استفاده از ویژگی‌های قسمتی از سخن (ارکان جمله) و وابستگی، تأثیر آنها محاسبه شد تا یکی دیگر از چالش‌های زبان فارسی پشت سر گذاشته شود.

پیشنهادها

یکی از کارهایی که موجب افزایش دقت پیش بینی در آنالیز احساس می شود، اضافه کردن یک ویژگی دیگر مانند تشخیص چگونگی خاتمه یافتن جمله است؛ به این معنا که جمله با کدام یک از علائم نگارشی (مانند نقطه، تعجب یا پرسش) تمام می شود. همچنین ممکن است جمله ای بدون به کار گرفتن لغت احساسی، مفهومی از حس نویسنده را دربرداشته باشد؛ درباره این موضوع نیز می توان تحقیق کرد.

منابع

علیمردانی، سعیده و آقایی، عبدالله (۲۰۱۵). اندیشه کاوی در زبان فارسی. فصلنامه مدیریت فناوری اطلاعات، ۲(۷)، ۳۶۲-۳۴۵.

References

- Ali Mardani, S. & Aghayi, A. (2015). Opinion Mining in Persian Language. *Journal of Information technology management*, 2(7), 345-362. (in Persian)
- Balahur, A. & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech and Language*, 28(1), 56-75.
- Banea, C. & Mihalcea, R. & Wiebe, J. (2014). Sense-level subjectivity in a multilingual setting. *Computer Speech and Language*, 28(1), 7-19.
- Barawi Hardyman, M. & Seng, Y. (2013). Evaluation of resource creations accuracy by using sentiment Analysis. *Procedia - Social and Behavioral Sciences*, 97(11), 522 - 527.
- Brychcín, T. & Konopík, M. (2014). Semantic spaces for improving language modeling. *Computer Speech and Language*, 28(1), 192 - 209.
- Dotti, F. (2013). Overcoming Problems in Automated Appraisal Recognition: the Attitude System in Inscribed Appraisal. *Procedia - Social and Behavioral Sciences*, 95(10), 442 - 446.
- Ghazi, D. & Inkpen, D. & Szpakowicz, S. (2014). Prior and contextual emotion of words in sentential context. *Computer Speech and Language*, 28(1), 76-92.
- Hadanoa, M. & Shimadaa, K. & Endoa, T. (2011). Aspect identification of sentiment sentences using a clustering Algorithm, *Procedia - Social and Behavioral Sciences*, 27(10), 22 - 31.

- Haddia, E. & Liua, X. & Shib, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science*, 17(5), 26 – 32.
- Lambova, D. & Paisa, S. & Diasa, G. (2011). Merged Agreement Algorithms for Domain Independent Sentiment Analysis. *Procedia - Social and Behavioral Sciences*, 27(10), 248 – 257.
- Devika, MD. & Sunitha, C. & Ganesha, A. (2016). Sentiment Analysis: A Comparative Study On Different Approaches, *Procedia Computer Science*, 87(5), 44 – 49.
- Yu, L. & Wu, Ch. & Jang, F. (2009). Psychiatric document retrieval using a discourse-aware model. *Artificial Intelligence*, 173(7), 817–829.