Journal of Information Technology Management
ISSN: 2008-5893
EISSN: 2423-5059
Summer 2018
Vol. 10, No. 2, PP. 375-389
DOI: 10.22059/jitm.2017.224726.1923

University of Tehran
Faculty of Management

Citation: Garshasbi, M., Rais-Rohani, A., & Kabaranzadeh Ghadim, M.R. (2018). Presenting a Text Mining Algorithm to Identify Emotion in Persian Corpus. *Journal of Information Technology Management,* 10(2), 375-389.

Presenting a Text Mining Algorithm to Identify Emotion in Persian Corpus

Masoud Garshasbi [1], Anahid Rais-Rohani [2], Mohammadreza Kabaranzadeh Ghadim [3]

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

مدیریت فناوری اطلاعات
دانشکدهٔ مدیریت دانشگاه تهران
دورهٔ ۱۰، شمارهٔ ۲
تابستان ۱۳۹۷
صص. ۳۸۹- ۳۷۵

Journal of Information Technology Management
DOI: 10.22059/jitm.2017.224726.1923

ارائهٔ الگوریتم متن‌کاوی به‌منظور تشخیص حس در متن‌های فارسی

مسعود گرشاسبی[۱]، آناهید رئیس روحانی[۲]، محمدرضا کاباران‌زاده قدیم[۳]

چکیده: در متن‌کاوی متن‌های فارسی، در زمینهٔ چگونگی استخراج ویژگی‌ها برای دسته‌بندی و بررسی نظرها در سایت‌های اجتماعی به‌منظور تشخیص قطبیت متن، مطالعاتی انجام شده است. هدف این پژوهش، ارائهٔ الگوریتمی برای آنالیز حس متن فارسی، بر اساس شش حس پایهٔ 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 تشدیدکننده‌ها و نفی‌کننده‌ها وجود دارند که روی حس متن تأثیر می‌گذارند. از این رو، الگوریتم به شش روش با در نظر گرفتن ویژگی‌های متفاوت نوشته شده است. در روش اول الگوریتم قابلیت تشخیص یک لغت عاطفی درون جمله را دارد؛ سپس قابلیت تشخیص تشدیدکننده، نفی‌کننده و لغات ایست اضافه می‌شود. نتایج به‌دست آمده از اجرای الگوریتم‌ها روی دو نمونه داده، نشان می‌دهد با در نظر گرفتن ویژگی‌های بیشتر، دقت الگوریتم نیز افزایش می‌یابد که در آن عاملِ قسمتی از سخن، بیشترین تأثیر را دارد.

واژه‌های کلیدی: *آنالیز احساس، اندیشه‌کاوی، داده‌کاوی، متن‌کاوی، وب‌کاوی.*

۱. مربی پژوهشی، پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران
۲. کارشناس ارشد مهندسی نرم‌افزار، دانشکدهٔ مکاترونیک دانشگاه آزاد اسلامی واحد کرج، تهران، ایران
۳. دانشیار گروه مدیریت، دانشکدهٔ مدیریت، دانشگاه آزاد اسلامی واحد تهران مرکزی، تهران، ایران

تاریخ دریافت مقاله: ۱۳۹۵/۱۰/۲۰
تاریخ پذیرش نهایی مقاله: ۱۳۹۶/۰۷/۰۲
نویسندهٔ مسئول مقاله: آناهید رئیس روحانی
E-mail: anahidrr@gmail.com

مقدمه

اندیشه‌کاوی یا آنالیز احساس زمینهٔ مطالعاتی است که سعی می‌کند احساس‌ها، رفتارها، نظرها و تحلیل افراد مختلف را نسبت به موجودیت‌ها و ویژگی‌های آن بیان کند. این موجودیت می‌تواند محصول، سرویس، سازمان، فرد و رخداد و موضوع باشد. هدف از آنالیز یا تحلیل احساسات، پیداکردن نظرهایی است که احساسی را نشان داده و قطبیت متن یا جهت‌گیری این نظرها را تشخیص می‌دهد.

افزایش اهمیت آنالیز احساسات با رشد رسانه‌های اجتماعی مانند نظرسنجی‌ها، فروم‌ها، انجمن‌های گفت‌وگو، وبلاگ‌ها، توییتر و شبکه‌های اجتماعی همزمان شده است. سیستم‌های آنالیز احساسات کمابیش در همهٔ زمینه‌های تجاری و اجتماعی به‌کار گرفته می‌شوند؛ زیرا نظرها و عقیده‌ها برای همهٔ فعالیت‌های انسانی مهم بوده و تأثیر شایان توجهی بر رفتار ما دارند. یکی از مهم‌ترین کاربرد آنالیز احساس در علوم روان‌شناسی است. در سال‌های اخیر تحقیقات بسیاری برای پیش‌بینی خودکشی از طریق پست‌های درج شده در شبکه‌های اجتماعی انجام شده است.

حال با توجه به مطالعات انجام شده، چالش‌هایی در این زمینه برای زبان فارسی وجود دارد که از جملهٔ آنها نقص پردازندهٔ زبان فارسی، صرف‌شدن افعال، استفاده از وندهای تصریفی و حجم زیادی از کلمات محاوره‌ای است. برای آنالیز می‌توان از پردازنده‌های زبان انگلیسی نیز استفاده کرد، اما ابتدا باید این کلمات ترجمه شوند و خود ترجمه روی دقت آنالیز تأثیر منفی می‌گذارد.

در پژوهش حاضر، از آنجا که داده‌های استفاده شده ساختاریافته نیستند، الگوریتم پیشنهاد شده مبتنی بر روش لغتنامه پیاده‌سازی می‌شود و به‌منظور افزایش دقت این روش، علاوه بر لغات عاطفی، از ویژگی‌های دیگری نظیر تشدیدکننده و نفی‌کننده نیز بهره برده شده است تا این الگوریتم بتواند حس متن را بر اساس شش حس پایه تشخیص دهد.

نوآوری این مقاله استفاده نکردن از الگوریتم‌های زبان دیگر است؛ بدین ترتیب متن با ترجمه دچار تغییر حس نمی‌شود. همچنین برای آنالیز، از تعداد تکرار لغت‌های مربوط به دستهٔ حسی (منظور از دستهٔ حس، شش حس پایهٔ خوشحالی، ناراحتی، خشم، ترس، نفرت و تعجب است) که به آن تعلق دارند، استفاده نشده است؛ بلکه در لغتنامه به لغت‌ها وزنی اختصاص داده شده و با محاسبهٔ نهایی آن و تأثیر مؤلفه‌های دیگر در جمله، حس آن مشخص شده است.

در ادامه، ابتدا پژوهش‌های گذشته مرور می‌شود؛ سپس چگونگی تهیهٔ لغتنامه‌ها و شرحی از مراحل اجرای الگوریتم پیشنهاد شده توضیح داده می‌شود و در آخر ضمن بحث و نتیجه‌گیری دربارهٔ اجرای الگوریتم روی متن‌های فارسی، مقایسه‌ای با الگوریتم نایوبیز صورت می‌گیرد.

پیشینهٔ پژوهش

آنالیز احساس که شاخه‌ای از متن‌کاوی محسوب می‌شود، در حال حاضر بـه موضـوع روز تبـدیل 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 الگوریتم‌هـا متمرکـز اسـت. در ادامه، توضیح کوتاهی از کارهای انجام شده با روش‌ها و دامنه‌های مختلف در ایـن زمینـه ارائـه می‌شود.

یو، وو و یانگ (۲۰۰۹) روی اسناد سایت مشـاوره دو مـدل بازیـابی مبتنـی بـر لغـت (مـدل و مدل OKAPI[2]) 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) به‌صورت دستی، جنبه‌ای از احساسات را در اسناد بازبینی یکی از بازی‌های استخراج شده از وب‌سایت شناسایی کردند. آنها در این تحقیق از روش خوشه‌بنـدی جملات و ابزار Bayon مبتنی بر الگوریتم دوبخشی استفاده کردند و به این نتیجـه رسـیدند کـه روش خوشه‌بندی بهتر از روش بدون خوشه‌بندی است.

لامبوا، پایشا و دیاسا (۲۰۱۱)، چهار نوع الگوریتم توافقی شـامل SAR[3]، MAA[4]، BMAA[5] و BMAADR[6] را ادغام کردند و نتیجه گرفتند که الگوریتم SAR بهترین نتیجه را تولید کرده است. روی مجموعه تفسیرهای دستیRIMDB، میـانگین دقـت الگـوریتم SAR حـدود ۷۷/۱ درصد بود. بهترین اجرا الگوریتم MAA دقتی به انـدازهٔ ۷۵/۶ داشـت، امـا ایـن میـزان دقـت از بهترین اجرای الگوریتم SAR بدتر بود. میـانگین دقـت الگـوریتم BMAADR نیـز حـدود ۸۰ درصد به‌دست آمد. به‌طور کلی، الگوریتم SAR اجرای بهتـری را در حالـت‌هـای مجموعـه داده منحصراً ذهنی و عینی دارد و الگوریتم BMAADR بهترین اجرای الگـوریتم بـرای مـتن‌هـای واقعی دارد است.

هادیا، هولیو و شیب (۲۰۱۳) به بررسی تأثیر پیش‌پـردازش روی مـتن‌هـای گرفتـه شـده از اینترنت پرداختند کـه نتیجـهٔ آن بـه افـزایش دقـت ماشـین بـردار منجـر شـد. بـرای داده‌هـای

---

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

پیش‌پردازش نشده، دقت در ماتریس تکرار ویژگی[۱] پیشرفت داشت و برای داده‌های پیش‌پردازش شده دقت هم در تکرار ویژگی و هم در حضور ویژگی[۲] افزایش یافت.

دوتی (۲۰۱۳) روی متن‌های سادهٔ 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 برچسب‌زن قسمتی از سخن[۳] استفاده کرد و موجب پیشرفت فراخوانی کلی شد.

باراوی هاردیمن و سنگ (۲۰۱۳) نیز روی ارزیابی دقت منابع، آزمایش‌هایی انجام دادند. روی ویدئوهای آموزشی و با اندازهٔ فاصلهٔ اقلیدسی برای مقایسهٔ نتایج به‌دست 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 می‌توان تفسیر کرد.

بالاهور و تورچی (۲۰۱۴) به آزمایش مقایسه‌ای با استفاده از یادگیری تحت نظارت و ماشین ترجمه به‌منظور تجزیه و تحلیل احساسات چندزبانه اقدام کردند. برای ترجمه از چهار سیستم SMT[۴]، گوگل، بینگ، moses و برای دسته‌بندی از ماشین بردار دسته‌ای[۵] استفاده کردند و نشان دادند استفاده از ماشین بردار دسته‌ای، تأثیر مثبتی بر نتایج می‌گذارد.

قاضی، اینکپن و اسپاکویچ (۲۰۱۴) با استفاده از روش لغتنامه به دسته‌بندی متن به شش حس پایه پرداختند. در این پژوهش سه مدل پایه، ماشین بردار و بهینه‌سازی حداقل پی در پی[۶] در نرم‌افزار وکا بررسی شده است. بر اساس نتایج به‌دست آمده، ماشین بردار روی دیتاست بزرگ نتیجهٔ بهتری داشت و LR[۷] در مقایسه با ماشین بردار روی ویژگی‌های مطرح شده در دسته‌ای از لغات بهتر جواب داد.

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

بریخن، کونوپیک (۲۰۱۴) الگوریتم‌های هال[۹]، کالز[۱۰]، بیگل[۱۱] و اصول و پارامترها[۱۲] را روی متن‌های اخباری به زبان‌های 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)

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

علیمردانی و آقایی (۲۰۱۵) روش الگوریتم ماشین بردار را با لغتنامه ترکیب کردند و آن را روی نظرهای ارائه شده در خصوص هتل با چهار فرضیهٔ تعداد تکرار کلمه‌ها، حضورداشتن و حضورنداشتن کلمه‌ها، حاصل ضرب تعداد تکرار در قطبیت و حاصل ضرب حضور داشتن و حضور نداشتن کلمه‌ها در قطبیت، به اجرا درآورند. نتایج نشان داد فرضیهٔ 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 برچسب خورده احتیاجی ندارد، بلکه به لغتنامهٔ 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 محاسبهٔ نتیجه وجود دارد. برای این کار به لغتنامهٔ فارسی نیاز است و چون در این زمینه لغتنامهٔ فارسی وجود نداشت، لغتنامه‌هایی در قالب فایل اکسل تهیه شد که در ادامه به شرح آن پرداخته می‌شود.

لغتنامه‌ها

لغتنامهٔ احساس: این لغتنامه شامل فهرست لغت‌هایی است که به یکی از شش حس پایهٔ بیان شده در پژوهش قاضی و همکارانش (۲۰۱۴) برچسب 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.

لغتنامهٔ تشدیدکننده: این لغتنامه شامل فهرست لغت‌های تشدیدکننده با برچسب 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. نمونهٔ لغتنامهٔ احساس

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

جدول ۲. نمونهٔ لغتنامهٔ تشدیدکننده

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

لغتنامهٔ نفی: این لغتنامه فهرستی از لغت‌هایی را دربرمی‌گیرد که نفی کننده‌اند و فعل یا پیشوندی با بار منفی دارند. در این لغتنامه هر لغت با برچسب v و n مشخص شده است. v به‌معنای فعل نفی یا پیشوند فعل نفی بوده و n پیشوند نفی اسم یا صفت است. جدول ۳ نمونه‌ای از لغتنامهٔ نفی را نشان می‌دهد.

جدول ۳. نمونهٔ لغتنامهٔ نفی

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

لغتنامهٔ 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، ضمایر، حروف اضافه و اسامی اشخاص و اشیا است.

جدول ٤. نمونهٔ لغتنامهٔ 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[1]. در این سامانه بخشی به نام آبشار وجود دارد که در آن دانشجویان می‌توانند در قالب متن با کارشناسان پشتیبانی، استادان و دانشجویان هم گروه خود گفت‌وگو کنند.

با آنالیز احساس این مکاتبات، می‌توان میزان رضایتمندی دانشجویان از استادان، محتواهای آموزشی و توقع دانشجو از سیستم مدیریت یادگیری الکترونیکی را ارزیابی کرد و به‌منظور ارتقای کیفیت آموزش الکترونیکی، افزایش خدمات مورد نیاز دانشجویان به‌صورت 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؛ به‌صورت دستی جمله‌های مکتوب شده به زبان فارسی جدا شدند. از لغتنامهٔ توقف نیز برای پاکسازی جمله‌ها از حروف اضافه و کلمات بی‌تأثیر استفاده شد.

1. Lms.iauec.ac.ir

ویژگی‌ها

انتخاب ویژگی در متن‌کاوی مرحلهٔ 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: ویژگی‌ها بر خود لغات تکیه دارند. این دسته برای جمله‌های ساده مناسب است. این لغات در لغتنامهٔ احساس وجود دارند و به هر یک برچسب حس زده شده است.
- قسمتی از سخن (ارکان جمله): لغاتی مانند تشدیدکننده و نفی‌کننده و همچنین افعال نفی هستند که در تشخیص حس جمله تأثیرگذارند. برای تشخیص آنها از لغتنامهٔ تشدیدکننده و نفی‌کننده استفاده شده است.
- وابستگی لغات: در جمله‌هایی که کلمهٔ تشدیدکننده روی حس لغت عاطفی تأثیر می‌گذارد یا پیشوندهای نفی‌کننده که روی لغات بعد از آن در جمله اثر دارد، در این دسته ویژگی‌ها قرار می‌گیرند. این ویژگی برای جمله‌هایی که بیش از یک لغت عاطفی دارد به‌کار می‌رود؛ زیرا ترتیب قرارگیری تشدیدکننده و نفی‌کننده مشخص می‌شود.
- تعداد لغات عاطفی در جمله: در جایی کاربرد دارد که جمله دربردارندهٔ چند لغت عاطفی با برچسب حس متفاوت است. در این ویژگی، بیشترین مجموع وزن لغات در یک دستهٔ مشابه، حس جمله را تعیین می‌کند.

یافته‌های پژوهش

پس از تهیهٔ لغتنامه‌های بیان‌شده در بخش قبل، الگوریتم به زبان سی شارپ با شش روش ترکیبی استفاده از ویژگی‌ها پیاده‌سازی شد و برای بررسی عملکرد آنها ویرایشگری تهیه گردید تا جمله را به‌صورت ورودی از کاربر گرفته و با انتخاب هر یک از ویژگی‌ها تشخیص حس توسط الگوریتم را ارزیابی کند.

روش نخست: استفاده از ویژگی لغت

الگوریتم تنها یک لغت عاطفی در جمله را با استفاده از لغتنامهٔ عاطفی تشخیص می‌دهد و حس آن لغت، حس جمله را بیان می‌کند. این روش برای جمله‌های 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 لغتنامهٔ عاطفی تشخیص داده می‌شوند، ولی این‌بار 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 تشدیدکنندهٔ 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 محاسبهٔ 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. نتیجهٔ این مرحله با وزن ماکزیمم جمله مقایسه

می‌شود، چنانچه مقدار بیشتری داشت، در وزن ماکزیمم جایگزین شده و حس آن به‌عنوان دستهٔ حس جمله در نظر گرفته می‌شود.

این کار تا پایان جمله ادامه می‌یابد و اگر از فعل نفی‌کننده استفاده شده باشد، حس کل جمله را تحت تأثیر قرار می‌دهد؛ بنابراین، الگوریتم حس به‌دست آمده را معکوس می‌کند. برای مثال، اگر حس جمله جزء دستهٔ خوشحالی بود، نفی آن جز دستهٔ ناراحتی می‌شود. برای دستهٔ حس‌هایی که معکوس وجود ندارد تا بتوان نفی آن حس را به یکی از دسته‌ها نسبت داد، یک دستهٔ دیگری به‌عنوان خنثی در نظر گرفته شده است که حس لغت در این دسته قرار داده می‌شود. 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 می‌شود. مانند روش چهارم نه‌تنها ترتیب قرار گرفتن لغت عاطفی با تشدیدکننده اهمیت دارد، بلکه ترتیب نفی‌کننده‌ها نیز مهم است. در لغتنامهٔ نفی‌کننده‌ها، برچسب زده شده روی هر یک مشخص می‌کند که لغت نفی‌کننده پیشوند است؛ برای مثال «بدون» یا فعل «نیست» که نفی‌کننده است.

جدول ۵. نمونه‌ای از جملات و نتایج به‌دست آمده در شش روش

جملات	روش ۱	روش ۲	روش ۳	روش ٤	روش ٥	روش ٦
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 دقیق‌تر بنویسید تست شد مشکلی نیست.	ناراحتی	ناراحتی	ناراحتی	ناراحتی	خوشحالی	خوشحالی
خیلی ممنونم از راهنمایییتون متأسفانه پشتیبان متوجه منظور من نشده	خوشحالی	ناراحتی	خوشحالی	خوشحالی	ناراحتی	ناراحتی

برای این کار متغیری گذاشته می‌شود تا لغت نفی‌کننده را همراه با برچسب آن ذخیره کند؛ سپس مانند حالت قبل، اگر لغت بعدی لغت عاطفی باشد و برچسب نفی‌کننده پیشوند باشد، آن را روی حس لغت عاطفی تأثیر می‌دهد 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 دستهٔ آن را جزء دستهٔ 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 برنامهٔ SQL Business Inteligence استفاده شد. دیتاسورس یا همان داده‌های آموزشی، شامل ۱۰۰ نمونهٔ داده‌های وب‌سایت اجتماعی آموزش الکترونیکی است که کاربران وارد کرده‌اند. روی این داده‌ها توسط الگوریتم روش ششم برچسب حس زدیم و نتیجهٔ آن توسط مفسر بررسی و تصحیح شد؛ سپس آن را روی داده‌های وب‌سایت اجتماعی آموزش دادیم. بقیهٔ پیاده‌سازی در برنامهٔ 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 نایوبیز 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 به‌کمک آن می‌توان به برنامهٔ 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 دامنهٔ متفاوت و همچنین با مقایسهٔ الگوریتم نایوبیز ارزیابی می‌شود.

نتایج به‌دست آمده از شش روش

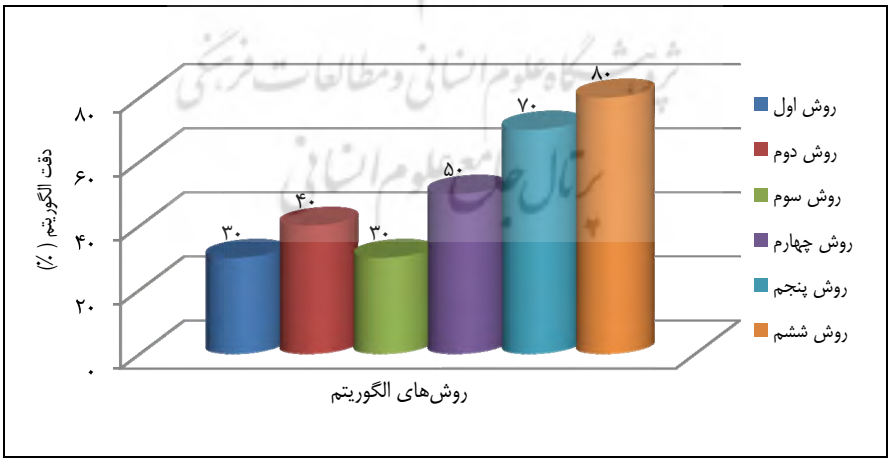
با گرفتن ایده از پژوهش هادیا و همکارانش (۲۰۱۳) که نشان دادند پیش‌پردازش تأثیر مثبتی بر دقت الگوریتم دارد و همچنین با بهره بردن از ویژگی‌هایی که در اجرای بهتر الگوریتم مقاله

قاضی و همکارانش (۲۰۱۴) مؤثرتر بودند، الگوریتم را بـه شـش روش ارزیـابی کـردیم و در هـر روش تعداد ویژگی‌ها را تغییر دادیم. حال به مقایسهٔ نتایج پرداخته می‌شود.

برای مقایسهٔ روش اول و دوم، جملهٔ «دیروز شاگردم را تنبیه کردم و الان پشیمان و ناراحـت هستم» بررسی شد. در روش اول که تنها لغات عـاطفی در نظـر گرفتـه مـی‌شـود، الگـوریتم بـا استخراج لغت تنبیه حس جمله را خشم تشخیص داد؛ اما در روش دوم که تعداد لغات عاطفی نیز در نظر گرفته می‌شود، تشخیص حس جمله، ناراحتی بود. دقت روش اول ۳۰ درصد و روش دوم ۴۰ درصد محاسبه شد که از تغییر درصد دقت می‌توان دریافت که ویژگی تعداد لغات عـاطفی در جمله بر دقت آنالیز حس توسط الگوریتم، تأثیر مثبتی می‌گذارد.

برای مقایسهٔ روش چهارم و پنجم، جملهٔ «خیلی ممنونم اما مشکلم حل نشد» برای آزمایش انتخاب شد. در روش چهارم که ویژگی «وابستگی» مطرح است و از ویژگی «قسمتی از جملـه» که ویژگی تشخیص فعل جمله است بهره نمی‌برد، 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 روش‌هـای قبلـی نشان می‌دهد ویژگی تعداد لغات عاطفی تأثیر بیشتری دارد؛ اما در مقایسـهٔ روش چهـارم و روش پنجم که دقت آن ۷۰ درصد بود، می‌توان گفت ویژگی قسمتی از سخن، به مراتب تأثیرگـذارتر از ویژگی وابستگی است.

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



شکل ۱. مقایسه روش‌های انجام شده روی داده‌های 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 نمونهٔ 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، دامنهٔ 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 نایوبیز مقایسه شد. این بار الگوریتم‌ها روی ۵۰ نمونهٔ دادهٔ 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 نایوبیز استفاده شده است، ۵۲ درصد حس جمله تشخیص درست و ۴۸ درصد جملات تشخیص نادرست داشت. برای بهبود روش یادگیری با الگوریتم نایوبیز، پیشنهاد می‌شود حجم داده‌های یادگیری بیشتر شود.

در این تحقیق ادعا شد که معنای اولیهٔ 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 نتیجهٔ 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 لغتنامهٔ 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 پیش‌بینی در آنالیز احساس می‌شود، اضافه‌کردن 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 می‌شود. همچنین ممکن است جمله‌ای بدون به‌کار گرفتن لغت احساسی، مفهومی از حس نویسنده را دربرداشته باشد؛ دربارهٔ این موضوع نیز می‌توان تحقیق کرد.

منابع

علیمردانی، سعیده و آقایی، عبداله (۲۰۱۵). اندیشه‌کاوی در زبان فارسی. *فصلنامهٔ 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 Appraisa. *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,s*, 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.