نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان – دانشکده مهندسی برق و کامپیوتر

آنالیز احساسی متون فارسی

شکوفه دشتبانی مشکانی[۱]، عبدالحمید پیله ور[۲]

آزمایشگاه مهندسی زبان، گروه کامپیوتر، دانشکده مهندسی، دانشگاه بوعلی سینا همدان

[۱]S.Dashtbani@basu.ac.ir , [2] pilevar@basu.ac.ir

چکیده:

هدف از عرضه این مقاله بررسی راهکار جدیدی به منظور آنالیز متون فارسی می باشد. در این مقاله آنالیز احساسی متون فارسی مورد بررسی قرار می گیرد و خصایص و ویژگی های روش پیشنهاد شده در این مقاله با روش های موجود در زبان انگلیسی مقایسه می گردد. اخیرا آنالیز احساسی متون در زبان های دیگر به ویژه زبان انگلیسی مورد مطالعه قرار گرفته است و روش هایی نیز به منظور آنالیز سریع متون انگلیسی ارائه شده است. روشی که در این مقاله مورد استفاده قرار گرفته است، تا حدودی نتیجه مشابه با روش های مورد استفاده در زبان انگلیسی داشته است. در این روش آنالیز احساسی با ساخت دسته های احساسی صورت می گیرد که روش ساخت دسته های احساسی بر اساس شباهت معنایی می باشد. استفاده از شباهت معنایی برای ساخت دسته های احساسی در زبان فارسی و آنالیز احساسی متون فارسی به طور کلی در حدود ۸۰ درصد موارد موفقیت آمیز بوده است.

کلمات کلیدی

هوش مصنوعی، شباهت معنایی[۱]، شباهت کلمات[۲]، دسته های احساسی[۳]

۱ – مقدمه:

پردازش زبان طبیعی یکی از شاخه های هوش مصنوعی می باشد که پردازش زبان فارسی در این مقاله مورد تحقیق قرار گرفته است. آنالیز احساسی ساختار متن از آنجا حائز اهمیت است که می تواند ابهامات را تا حدودی کاهش دهد. تحقیقات گسترده ای در زمینه آنالیز احساسی متون انجام شده است ، استفاده ازآنالیز احساسی به منظور استخراج اندیشه[4] [1] و همچنین طبقه بندی و آنالیز احساسی سخنان گفته شده[5] [2] از قبیل کارهای انجام شده در این زمینه می باشد.

یکی از تکنیک های اساسی که در این مقاله مورد استفاده قرار گرفته است ، استفاده از شباهت کلمات در زبان فارسی می باشد که از آن برای پیدا کردن شباهت لغات متن فارسی با دسته های احساسی مورد استفاده قرار گرفته است. راهکارهایی تا کنون برای زبان انگلیسی ارائه گردیده است ، از قبیل بدست آوردن شباهت با استفاده از بردار ویژگی[6] [3] ، تخمین بر اساس فاصله بین دو نود[7] [4] ، بر اساس ارتباط مفهومی[8] [5] ، استفاده از منطق فازی [6] و همچنین شباهت معنایی را می توان بر اساس محاسبات آماری[9] [7] نیز محاسبه کرد. ساختن دسته های احساسی نیازمند مطالعه ی دقیق در دستور زبان فارسی و یا استفاده از یک متخصص در زمینه ادبیات فارسی می باشد که در این مقاله 40 دسته احساسی در زبان فارسی با استفاده از کتب دستور زبان فارسی [8] معرفی می شود که از این 40 دسته احساسی پس از تجزیه متن ورودی به کلمات تشکیل دهده اش ، برای آنالیز متن استفاده می گردد.

ابهامات متن با توجه به اینکه لغت ورودی در چه دسته های احساسی می تواند قرار بگیرد، مدیریت خواهند شد. رفع ابهام از این جهت است که علاوه بر بیان احساس ، کلمه در زبان طبیعی را نیز بیان می کند[9].علاوه بر طبقه بندی معنایی ، در اینجا از چگالی عددی کلمات در متن ورودی نیز برای مدیریت ابهامات استفاده می شود. تمامی کلمات احساسی متن به همراه چگالی عددی و عددی که نشان دهنده میزان شباهت کلمه ورودی به دسته های احساسی می باشد را در یک فایل پیوست به متن ورودی قرار می دهیم، در نهایت با توجه به آن متن ورودی را مورد آنالیز قرار می دهیم. در زبان انگلیسی اخیرا یکسری محاسبات احساسی نیز ارائه گردیده است[10]. در آخر هم پیشنهاد هایی در

راستای انتخاب راهکارهای مناسب به منظور تسریع در روش پیشنهاد شده، ارائه شده است.

۲- ساخت دسته های احساسی:

همسانی، همانندی و یا شباهت مفهومی است که اخیرا به طور گسترده مورد استفاده قرار گرفته است و تعاریف گوناگونی برای آن ارائه گردیده است. تعریفی که در این مقاله ارائه شده است ، تلاش شده است که یک کاربرد خاص از دانش ارائه شده توسط آن ، ارائه گردد. همسانی در اینجا منظور همانندی در احساس می باشد که یک دانش از مفهوم را در اختیار ما قرار می دهد. ما در اینجا مدلی برای بدست آوردن شباهت ارائه می دهیم، مدلی که در اینجا برای بدست آوردن شباهت مورد استفاده قرار می گیرد یک مدل مبتنی بر احتمال است.

۲-۱- روش پیشنهادی:

در زبان انگلیسی چندین روش برای بدست آوردن شباهت ارائه شده است، برای مثال برای تعیین شبیه بودن دو کلمه می توان ریشه دو کلمه را در نظر گرفت ، اگر ریشه سه تایی [10] دو کلمه در زبان انگلیسی یکسان باشد آن دو کلمه با هم شبیه هستند. اما در زبان فارسی این روش امکان پذیر نسیت زیرا به طور مثال علم و علوم دارای یک ریشه هستند اما ریشه سه تایی آنها باهم یکسان نیست. در بین روش های موجود روشی که برای زبان فارسی قابل استفاده است روش شباهت معنایی [5] می باشد که در نرم افزار wordnet هم از آن استفاده شده است. در این مقاله برای ساخت دسته های احساسی با اندکی تغییرات از این روش استفاده شده است. برای شروع به یک لغت نامه جامع از کلمه هایی که بار احساسی دارند و یا می توانند در جمله بار احساسی داشته باشند، نیاز داریم. لغت نامه ای که ما جمع آوری کردیم مشتمل بر حدود هزار کلمه در زبان فارسی است که یا خود کلمه به تنهایی بار احساسی دارد و یا اینکه به تنهایی احساسی ندارد اما درمتن می تواند بار احساسی بگیرد . گردآوری لغات احساسی باید تحت نظارت متخصصین در زمینه ادبیات باشد، هرچقدر که لغت نامه کامل تر باشد ، دقت کار افزایش می یابد. مرحله بعدی ساختن درخت از تمامی لغات موجود در لغت نامه می باشد ، درختی که در اینجا ارائه شده است مبتنی بر لغات احساسی زبان فارسی می باشد. نحوه ساخت درخت به این گونه است که نود ریشه شامل تمام کلمات موجود در لغت نامه می باشد . انشعاب از نود ریشه براین اساس که کدامیک از لغات می توانند در یک دسته قرار بگیرند، انجام می شود. درخت ساخته شده در این تحقیقات شامل چهار سطح است که سطح آخر درخت تشکیل دهنده دسته های احساسی می باشد. در واقع هر برگ درخت شامل کلماتی است که از نظر احساسی به هم شباهت دارند. بعضی از طبقه ها یا برگ ها هم چیزهایی که به طور ضمنی دارای احساس هستند و مستقیما احساسی را بیان نمی کنند ، نگهداری می کنند. قدرت درخت احساسی که می سازیم به خوشه بندی[11] صحیح کلمه های لغت نامه می باشد. اشاره شد که این طبقه بندی کلمات باعث مدیریت ابهام کلمه های متن خواهد شد. در این مقاله ۲۰ دسته احساسی در نظر گرفته شده است که یک کلمه می تواند در برگ های دیگر درخت هم تکرار شده باشد، به عنوان مثال مرگ کلمه ای است که می تواند هم احساس رنج و ناراحتی را داشته باشد و هم احساس خوف وترس.

برچسبی که نشان دهنده احساس یک دسته است به این صورت انتخاب می گردد که از بین تمام کلماتی که متعلق به یک دسته هستند، آن کلمه ای که بیشترین شدت را در بیان احساس آن دسته دارد، به عنوان برچسب آن دسته در نظر گرفته می شود.

۳- آزمایشات انجام شده:

۳-۱- محاسبه عدد شباهت:

پس از پیاده سازی درخت ، برای آنالیز کردن نیاز است تا تمام کلمه های متن ورودی را مورد بررسی قرار دهیم به همین منظورنیاز است تا برای هر عدد ورودی، ما عدد شباهت آن با دسته های احساسی را محاسبه می کنیم.

برای هر متنی که از ورودی دریافت می کنیم ما یک جدول می سازیم که تمام کلمات متن (به جز حروف اضافه، حروف عطف و سایر حروفی که فاقد احساس هستند) در این جدول قرار می گیرند .در این جدول ۲۲ ستون وجود دارد که به ازای هر کدام از این کلمات ما درخت را جستجو می کنیم. جستجوی درخت از برگ ها که دسته های احساسی هستند، آغاز می گردد و به طرف بالا پیش می رود. پس ما میزان شباهت هر کلمه با هر ۲۰ دسته احساسی را محاسبه می

کنیم. برای این منظور ما برچسب یک دسته احساسی که در آن دسته وجود دارد و کلمه ای است که بیشترین شدت را در بین کلمه های آن دسته را دارد، انتخاب می کنیم .
بنابراین عدد شباهت عددی است که میزان شباهت یک کلمه از متن ورودی را با برچسب دسته احساسی را نشان می دهد.
عدد شباهت از طریق فرمول۱ محاسبه می گردد:
اگر کلمه X عضو دسته C باشد و اگر برچسب دسته مورد بررسی X' باشد و خود X' در دسته C' باشد و فرض می شود که X و X' هر دو از هم مستقل باشند.

$$Sim(X,X')=\frac{2\log p(C'')}{\log p(C)+\log p(C')} \qquad (1)$$

$P(c)$ احتمال انتخاب یک کلمه به طور تصادفی در کلاس c می باشد.
$P(c')$ احتمال انتخاب یک کلمه به طور تصادفی در کلاس c' می باشد.
c'' کلاسی است که هم کلمه X و هم کلمه X' به آن تعلق دارد یعنی همان کلاس پدر c و c' و $p(c'')$ احتمال انتخاب یک کلمه به طور تصادفی در کلاس c'' می باشد.
عدد شباهت عددی در بازه $[0,1]$ می باشد، که در صورتی که هر دو کلمه x و x' در یک دسته باشند، این عدد ۱ خواهد بود. در فرمول۲ به عنوان مثال عدد شباهت دو کلمه "درد" و "شادی" به صورت زیر محاسبه می گردد:

$$Sim(\text{درد,شادی}) = \frac{2\log p(\text{احساسی})}{\log p(\text{شادی})+\log p(\text{درد})} \qquad (2)$$

عدد شباهت حاصل از این معادله ۰,۰۵۲۸ می شود که همان طور که ملاحظه می کنید یک عدد فوق العاده کوچکی می باشد و به این میزان شباهت هم به این دلیل است که هر دو کلمه مستقیما بیان گر یک حس هستند.

مثال دیگر که دارای شباهت زیادی هستند عبارت است از شباهت دو کلمه "رضایت " و "شادی "(فرمول۳).

$$Sim(\text{رضایت,شادی}) = \frac{2\log p(\text{خوب احساس})}{\log p(\text{شادی})+\log p(\text{رضایت})} \qquad (3)$$

عدد شباهت این دو کلمه ۰,۷ می باشد که ملاحظه می کنید که میزان بالایی دارد .

به همین ترتیب در پایان ما جدولی از لغات متن ورودی به همراه عدد شباهت آنها به ۲۰ دسته احساسی خواهیم داشت.

به عنوان نمونه شباهت ۸ کلمه را با دسته احساسی علاقه در زیر مشاهده می کنید:

جدول ۱. شباهت ۸ کلمه احساسی با دسته علاقه

شباهت	کلمه ورودی
۰	خسته
۰,۰۲	رضایت
۰	دیوانگی
۰	غم
۰	مرگ
۱	عشق
۰	درد

۳-۲- محاسبه چگالی:

پس از جدا کردن کلمات متن و محاسبه عدد شباهت ، یک ستون دیگر در این جدول وجود دارد که چگالی کلمه ی احساسی در متن ورودی می باشد. در واقع عدد چگالی به این دلیل محاسبه می گردد که در آنالیز احساسی متن ممکن است که عدد شباهت یک کلمه به دو دسته تقریبا یکسان باشد ، در این صورت چگالی هر کدام از کلمات در اینجا نقش تعیین کننده ای خواهد داشت. به عنوان مثال کلمه " لذت " دارای عدد شباهت یکسان نسبت به دسته های "رضایت" و "شادی" می باشد. بنابراین هرچقدر که کلمات دسته "رضایت" درمتن ورودی بیشتر تکرار شده باشند نسبت به کلمات دسته "شادی" ، در نهایت احساس متن "احساس رضایت " خواهد بود.
مدیریت ابهامات که در ابتدای مقاله اشاره شد ، در اینجا نمایانگر می شود . به عبارت دیگر برای یک کلمه ورودی

که می تواند عضو چندین دسته احساسی باشد، وقتی که در نهایت با توجه به عدد شباهت و عدد چگالی تعیین می کنیم کلمه متعلق به کدام دسته احساسی می باشد، ابهامی که در مفهوم آن وجود داشته است را از بین می بریم.

۴-نتایج حاصل:

پس از اینکه ما برای هر متن ورودی جدول مذکور را ساختیم، نوبت به تعیین دسته احساسی بر اساس جدول می باشد. ما با توجه به عدد شباهت هر ستون و چگالی میزان درصد تعلق به هر دسته را محاسبه می کنیم و در نهایت هر دسته که دارای درصد بالاتری بود ، احساس غالب متن را نشان خواهد داد.

نتایج را برای متن زیر ، به عنوان یک متن نمونه ، مشاهده می کنید:

" آیا گمان می کنید که خوشبخت بودن آرزویی دور و دست نیافتنی است؟ آیا معتقدید که حوادث خارجی، خوشبختی را به شما هدیه می کنند؟ اگر گمان می کنید که خوشبختی در اندیشه و ذهن شما پنهان شده است ، فرصت برای ساختن یك زندگی مملو از شادی و رضایتمندی در پیش روی شما قرار دارد ، حق شناسی و قدردانی لازمه خوشبختی است . پس بهتر است که هر روز برای کمترین چیزی که در اختیار دارید، سپاسگزار باشید."

جدول ۲ .عدد شباهت کلمات ورودی برای ۵ دسته احساسی

	علاقه	رضایت	آشفتگی	هیجان	شادی
خوشبخت	0.27834463	۱	۰	0.2883142	۰,۵۵۸۷۱۴۲
آرزو	0.27834463	۱	0.054498509	0.2883142	0.28831420
حوادث	0.05135125	0.057034928	0.27834463	0.05308	۰
هدیه	0.27834463	۱	۰	0.2883142	0.28831420
شادی	0.25984831	۱	۰	۰,۸۸۴۲۲	۱
رضایتمندی	0.27834463	۱	۰	0.28831420	۰,۵۵۸۷۱۴۲
حق شناسی	0.2783446328	۱	۰	0.2883142008	۰,۵۵۸۷۱۴۲
قدردانی	0.2783446328	۱	۰	0.2883142008	۰,۵۵۸۷۱۴۲
سپاسگزار	0.2783446328	۱	۰	0.2883142008	۰,۵۵۸۷۱۴۲

جدول ۳. آنالیز احساسی متن برای ۵ دسته احساسی

نتیجه	علاقه	رضایت	آشفتگی	هیجان	شادی
درصد	%۱۷	%۵۳	%۲۵	%۲۵	%۲۵

نتایج موجود در جدول شماره ۳ علاوه بر عدد شباهت موجود در جدول ۲ ، بر اساس چگالی کلمات ورودی در متن نیز محاسبه می گردند.

همانطور که مشاهده کردید نتایج به صورت درصدی میزان تعلق متن را به هر کدام از دسته های احساسی بیان کرده است و احساس غالب آن احساس رضایت می باشد که با مطالعه متن نیز می توان این احساس را به صورت شهودی درک کرد.

همان طور که در روش های فازی انجام شده برای زبان انگلیسی [6] به صورت نموداری میزان تعلق متن به هر کدام از دسته های فازی نشان داده شده است، در تحقیقات آینده، نیز تلاش برای ترکیب روش های فازی با روش ارایه شده نیز انجام خواهد شد. استفاده

از روش های فازی ارائه میزان تعلق متن به دسته های احساسی را قابل فهم تر و کاراتر خواهد کرد.

۵. نتیجه گیری:

استفاده از روش شباهت معنایی ابهامات را خوب مدیریت می کند. این روش علاوه براینکه دقت

بالایی دارد نسبت به روشهای مشابه سرعت بسیار بالاتری دارد. در هر صورت تلاش هایی که در این مقاله گزارش شده است، می تواند مقدمه ای برای مطالعات و تحقیقات بیشتری در زبان فارسی باشد.

برنامه ای که برای آینده در راستای این تحقیقات انجام خواهیم داد، عبارتند از:

- گسترش دسته های احساسی به یک سری زیر دسته ها، به منظور افزایش توانایی بیان جزئیات احساسی بیشتر در متن ورودی
- استفاده از حوزه های دیگر هوش مصنوعی برای افزایش دقت در تعیین دسته های احساسی

۶. منابع:

1. G.Grefenstette, Yan Qu, J. G. Shanahan, D. A. Evans, "**Coupling Niche Browsers and Affect Analysis for an Opinion Mining Application**", Gegory Clairvoyance Corporation, 5001 Baum Bd, Suite 700, Pittsburgh, PA, 15213-1854, USA

2. Deb Roy and Alex Pentland, "**Automatic Spoken Affect Analysis and Classification**", MIT Media Laboratory, Perceptual Computing Group, 20 Ames St. Cambridge, MA 02129 USA.

3. Dekang Lin, "**An Information-Theoretic Definition of Similarity**", University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2

4. Jay J. Jiang and David W. Conrath, "**Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy**", 1997, Taiwan

5. [Resnik, 1995b] Resnik, P. (1995b). "**Using information content to evaluate semantic similarity in a taxonomy**", In *Proceedings of IJCAI-95*, pages 448–453, Montreal, Canada.

6. Pero Subasic and Alison Huettner ,"**Affect analysis of text using fuzzy semantic typing**", CLARITECH Corporation, Justsystem Group, 5301 Fifth Avenue, Pittsburgh, PA 15232, USA

7. Jay J. Jiang, "**Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy**", Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1, David W. Conrath, MGD School of Business, McMaster University, Hamilton, Ontario, Canada L8S 4M4.

۸. دکتر حسن احمدی گیوی و دکتر حسن انوری ، "**دستور زبان فارسی** " .

9. Lotfi A. Zadeh, "***Fuzzy Logic Computing with Words*** " , IEEE Transactions on Fuzzy Systems, 2, 103-111, 1996.

10. Rosalind W. Picard, "***Affective Computing***", MIT Press, 1997.

---

(۱). Semantic similarity
(۲). Word Similarity
(۳). Affect category
(۴). Analysis for an Opinion Mining Application
(۵). Automatic Spoken Affect Analysis and Classification
(۶). Feature vector
(۷). Edge-based (Distance) Approach
(۸). Information concept
(۹). Semantic Similarity Based on Corpus Statistics and Lexical
(۱۰). Trigram
(۱۱) Clustering