حس‌نگار: شبکهٔ واژگان حسی فارسی

احسان عسکریان*[۱]، محسن کاهانی[۲] و شهلا شریفی[۳]

[۱و۲] آزمایشگاه فناوری وب، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران

[۳] گروه آموزشی زبان‌شناسی همگانی، دانشکده ادبیات و علوم انسانی، دانشگاه فردوسی مشهد، مشهد، ایران

چکیده

وظیفهٔ اصلی نظرکاوی استخراج و تشخیص حس مثبت یا منفی (رضایت‌مندی) افراد، از روی اطلاعات متنی است. نبود یک واژه‌نامهٔ حسی فارسی عامل یکی از چالش‌های اصلی نظرکاوی در زبان فارسی است. در این مقاله روشی جدید برای تولید شبکهٔ واژگان حسی فارسی (حس‌نگار) با استفاده از منابع زبانی فارسی و انگلیسی ارائه می‌شود. همچنین پیکرهٔ نظرات فارسی ایجادشده برای انجام پژوهش‌های نظرکاوی، معرفی خواهندشد. برای تولید حس‌نگار ابتدا شبکهٔ واژگان جامع زبان فارسی (فردوس‌نت) ساخته شده‌است؛ سپس میزان حس هر گروه هم‌معنی در شبکهٔ واژگان حسی انگلیسی به کلمات متناظر آنها در حس‌نگار (شبکه واژگان حسی فارسی) نگاشت می‌شود. در آزمایش‌های انجام‌شده، مشخص شد که حس‌نگار دارای دقت ۰/۸۶ و نرخ بازیابی ۰/۷۵ است و می‌تواند به‌عنوان واژه‌نامهٔ حسی مرجع برای زبان فارسی استفاده شود.

واژگان کلیدی: نظرکاوی، شبکهٔ واژگان فردوس‌نت، واژه‌نامهٔ حسی، ابزارهای پردازش متون زبان فارسی.

HesNegar: Persian Sentiment WordNet

Ehsan Asgarian*[1], Mohsen Kahani[2] & Shahla Sharifi[3]

[1,2]Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

[3]Department of Linguistics, Faculty of Letters and Humanities, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

Awareness of others' opinions plays a crucial role in the decision making process performed by simple customers to top-level executives of manufacturing companies and various organizations. Today, with the advent of Web 2.0 and the expansion of social networks, a vast number of texts related to people's opinions have been created. However, exploring the enormous amount of documents, various opinion sources and opposing opinions about an entity have made the process of extracting and analyzing opinions very difficult. Hence, there is a need for methods to explore and summarize the existing opinions. Accordingly, there has recently been a new trend in natural language processing science called "opinion mining". The main purpose of opinion mining is to extract and detect people's positive or negative sentiments (sense of satisfaction) from text reviews. The absence of a comprehensive Persian sentiment lexicon is one of the main challenges of opinion mining in Persian.

In this paper, a new methodology for developing Persian Sentiment WordNet (HesNegar) is presented using various Persian and English resources. A corpus of Persian reviews developed for opinion mining studies are introduced. To develop HesNegar, a comprehensive Persian WordNet (FerdowsNet), with high recall and proper precision (based on Princeton WordNet), was first created. Then, the polarity of each synset in English SentiWordNet is mapped to the corresponding words in HesNegar. In the conducted tests, it was found that HesNegar has a precision score of 0.86 a recall score of 0.75 and it can be used as a comprehensive Persian

* Corresponding author

* نویسندهٔ عهده‌دار مکاتبات

SentiWordNet. The findings and developments made in this study could prove useful in the advancement of opinion mining research in Persian and other similar languages, such as Urdu and Arabic.

Keywords: Opinion Mining, FerdowsNet (Persian WordNet), Sentiment Lexicon, Persian Text Processing Tools

۱- مقدمه

آگاهی از تجربیات، نظرات و دیدگاه افراد نقش اساسی در فرایند تصمیم‌گیری مشتریان ساده تا مدیران سطح بالای شرکت‌های تولیدکننده و سازمان‌های مختلف دارد. امروزه با پیدایش تارنمای ۲/۰، حجم بسیاری از متون مربوط به نظرات افراد ایجاد شده‌است. ولی کاوش در حجم انبوه مستندات، منابع نظرسنجی مختلف و وجود نظرات مغایر دربارهٔ یک موجودیّت، فرایند استخراج و پردازش نظرات را بسیار دشوار ساخته‌است. بنابراین نیاز به روش‌هایی برای کاوش و تحلیل نظرات موجود در وب احساس می‌شود. بدین منظور در دههٔ اخیر گرایش جدیدی در علم پردازش زبان‌های طبیعی[1] به نام نظر کاوی[2] ایجاد شده‌است. یکی از مهمترین وظایف این حوزه دسته‌بندی حسی مستندات بر اساس بار حسی[3] مثبت یا منفی (میزان رضایت‌مندی) آنهاست.

اغلب پژوهش‌های نخستین در زمینهٔ نظرکاوی، سعی در دسته‌بندی نظر یا حس کلی یک متن، در قالب دو دستهٔ حس مثبت و منفی، داشتند [1]. در ادامه، پژوهشگران سعی در تعیین درجهٔ رضایت (میزان کمّی در بازه مشخص) متن نظرات (به جای دسته‌بندی قطبی یا دوحالته) کردند [2]. برای انجام این دسته‌بندی‌ها اغلب از روش‌های نظارت‌شده[4] که برچسب نمونه‌ها در آنها به‌صورت دستی مشخص شده‌بود، در حوزه‌های محصولات تجاری که نظرات به‌تصریح بیان شده‌بودند، استفاده می‌کردند. مشکل اصلی تحلیل حس در واحد متن، فرض یکسان بودن مضمون در همهٔ متن یا متون جمع‌آوری‌شده است؛ درحالی‌که ممکن است بخش‌های مختلف متن (یا متون مختلف)، دارای مضمون‌های متفاوتی باشند. بنابراین قبل از تحلیل حس، لازم است مضمون بخش‌های مختلف شناسایی و جدا از هم بررسی شوند. درنتیجه، پژوهش‌گران نظرکاوی کار تحلیل حس را در واحد جمله[5] [3] یا قاب‌های معنایی[6] [4]، ادامه دادند. معمولاً در این واحد، از تحلیل ذهنیت[7] برای تمیز دادن جملات حاوی حس از جملات فاقد حس (شامل واقعیت مانند اخبار) استفاده می‌شود.

در سال‌های اخیر بسیاری از پژوهش‌های صورت گرفته در این زمینه به سمت زبان‌های غیرانگلیسی (به‌ویژه اسپانیایی، چینی، آلمانی، چکوسلواکی[8] و عربی)، معطوف شده‌است [5-8]. البته روش‌های نسبتاً جدیدی برای نظرکاوی چندزبانه[9] در حال ایجاد هستند. در زمینهٔ نظرکاوی چندزبانه بیشتر کارهای انجام‌شده به تحلیل حس در واحد متن (سند) و یا تحلیل ذهنیت و حس در واحد جمله پرداختند [9,10]. غالباً این روش‌ها با ایده‌های به‌نسبت ساده‌ای از مترجم‌های ماشینی به منظور بهره‌گیری از مجموعه اصطلاحات حاوی حس و سایر منابع و ابزارهای زبان انگلیسی برای زبان مورد نظر خود، استفاده می‌کنند [11-14]. ولی با توجّه به تفاوت قواعد نحوی زبان‌ها، اصطلاحات حاوی حس و سایر پیچیدگی‌های زبانی، نتایج به‌دست‌آمده از این دسته روش‌ها دقت مناسبی برای استفاده در زبان‌های مختلف ندارند. در حال حاضر پژوهش‌های بسیار اندکی بر روی نظر کاوی در زبان فارسی صورت گرفته‌است. البته کمبود منابع (از قبیل پیکره استاندارد متون نظرات و مجموعه واژگان حاوی حس استاندارد) و ابزارهای استاندارد پردازش متن مناسب و در دسترس و سایر پیچیدگی‌های نحوی و دستوری زبان فارسی مانع بزرگی برای انجام بررسی‌های نظرکاوی در زبان فارسی است.

ابزارهای پردازش زبان طبیعی و شبکهٔ واژگان[10] (هستان‌شناسی ریخت‌شناسی بین کلمات[11]) نقش مهمی در کاربردهای پردازش زبان طبیعی و بازیابی اطلاعات متنی

[1] Natural Language Processing

[2] Opinion Mining

[3] polarity

[4] supervised

[5] Sentence level sentiment analysis

[6] Semantic frame

[7] Subjectivity Analysis

[8] Czech

[9] Multi-Lingual

[10] WordNet

[11] lexical ontology of words

موجود در وب دارد. از این‌رو در این پژوهش، ابتدا مجموعه‌ای از ابزارهای پردازش متون زبان فارسی مورد نیاز ایجاد شده و شبکهٔ واژگان جامع زبان فارسی (فردوس‌نت) معرفی می‌شوند. سپس با استفاده از شبکهٔ واژگان ایجادشده، مفاهیم و عبارات حسی استخراج‌شده برای زبان انگلیسی به زبان فارسی نگاشت داده‌می‌شوند. تهیهٔ واژه‌نامهٔ عبارات حاوی حس[1] برای زبان طبیعی یکی از حوزه‌های پژوهشی در نظرکاوی به‌شمار می‌رود. علاوه‌براین، جهت ارزیابی روش‌های دسته‌بندی احساسات، پیکرهٔ متون نظرات مربوط به محصولات مختلف تجاری در زبان فارسی جمع‌آوری و بخشی از آن توسط انسان حاشیه‌نویسی (برچسب‌گذاری) حسی شده‌است.

در بخش بعدی مقاله به مرور روش‌های دسته‌بندی احساسات پرداخته می‌شود. در بخش سوم ابزارهای ایجادشده برای پردازش متون زبان فارسی معرفی و به‌اختصار توضیح داده خواهند شد. در بخش چهارم، نحوهٔ ایجاد شبکهٔ واژگان فردوس‌نت و تولید شبکهٔ واژگان حسی فارسی (حس‌نگار) توضیح داده خواهد شد. در بخش آخر نیز نتایج ارزیابی کیفیت ابزارهای پردازش متن مختلف ایجادشده، شبکه‌های واژگان مختلف زبان فارسی و حس‌نگار گزارش خواهند شد.

۲- مروری بر کارهای گذشته

در چند سال گذشته، ساخت مجموعهٔ لغات حاوی حس با بار حسی مثبت و منفی، یکی از روش‌های مورد توجه پژوهش‌گران برای تشخیص حس جملات بوده‌است. به‌طور کلی روش‌های تحلیل احساسات را می‌توان به دو گروه ۱- روش‌های مبتنی بر واژه‌نامهٔ حسی و استفاده از دانش زمینه (یادگیری بدون ناظر یا شبه‌ناظر) و ۲- روش‌های یادگیری باناظر تقسیم‌بندی کرد. دقت روش‌های مبتنی بر واژه‌نامهٔ حسی به‌طور کامل وابسته به مجموعهٔ لغات حاوی حس و وزن‌های از پیش‌تعیین شده است [15]. این روش‌ها به‌طور بدون ناظر و برای حوزه‌های عمومی قابل‌استفاده هستند [16]. در رویکرد دوم (دسته‌بندی حسی متون) نیز از واژگان حسی به‌عنوان یکی از ویژگی‌های مهم متن نظرات استفاده می‌شود.

از دیگر روش‌های تشخیص حس عبارات، استفاده از روش‌های محاسبهٔ شباهت معنایی کلمات است. در این روش‌ها برای تشخیص حس نظرات به‌طورمعمول از شباهت معنایی عبارات و فهرست کوچکی ازکلمات حاوی حس نخستین استفاده می‌شود. برای محاسبهٔ شباهت معنایی به‌طورمعمول از روش‌های زیر استفاده می‌شود:

۱- مبتنی بر شبکهٔ واژگان یا سایر لغت‌نامه‌ها و دانش‌نامه‌ها [17]

۲- روابط وابستگی نحوی بین عبارت حاوی حس با کلمات موجود در واژه‌نامهٔ حسی

۳- هم‌رخدادی عبارات حاوی حس با کلمات موجود در فهرست نخستین کلمات حاوی حس (روش‌های یادگیر بدون ناظر) در درون پیکره‌های مختلف مستندات[2] [18].

این رویکرد را زیرمجموعه‌ای از رویکرد تشخیصی حس مبتنی بر مجموعه لغات می‌توان به‌شمار آورد؛ با این تفاوت که فهرست کلمات حاوی حس برای مستندات ورودی (داده شده) تشکیل می‌شود و از این‌رو کارکرد بهتری برای تشخیص حس عبارات در حوزه‌های مختلف خواهدداشت.

با توجه به وابستگی زیاد روش‌های مختلف تحلیل احساسات به لغت‌نامهٔ واژگان حسی، در این بخش به توضیح روش‌های مختلف واژه‌نامه حسی پرداخته می‌شود؛ به‌طور کلی، از سه رویکرد زیر برای تولید واژه‌نامه‌های حسی استفاده می‌شود:

۱- مبتنی بر پیکره [1,19-21]

۲- مبتنی بر لغت‌نامه و پایگاه دانش [17,22-24]

۳- مبتنی بر روش‌های یادگیر باناظر [25-27]،

در رویکرد ساخت لغت‌نامهٔ حسی با استفاده از روش‌های یادگیر باناظر، به داده‌های آموزشی دارای برچسب حسی اولیه نیاز است. به‌دلیل مشکلات برچسب‌گذاری حسی کلمات (تهیه پیکرهٔ آموزشی) گاهی از این رویکرد برای استخراج واژگان حسی در دامنهٔ محدود (خاص) نظرات استفاده می‌شود. اغلب از ترکیب این رویکرد با روش‌های مبتنی بر لغت‌نامه (پایگاه دانش) و یا پیکرهٔ نظرات (با قالب مشخص) استفاده می‌شود.

روش‌های مبتنی بر پیکره، از پیکره‌های متنی به‌نسبت بزرگ و از قواعد زبان‌شناسی استفاده می‌کنند. به‌طورمعمول برای ایجاد واژه‌نامهٔ حسی وابسته به دامنه (موضوع) خاص استفاده می‌شود. البته با درنظرگرفتن پیکره‌های متنی بزرگ از این رویکرد برای تولید واژه‌نامه‌های حسی عمومی (مستقل از دامنه) نیز می‌توان استفاده کرد. هاتزیواسیلوقلو و همکارش [19] از قواعد زبان‌شناسی برای تعیین بار حسی صفت‌ها

[1] Sentiment lexicon

[2] مستندات نظرات کاربران، صفحات وب یا نتایج خروجی موتورهای جستجو

استفاده کردند. بدین منظور، از حروف ربط[1] بین صفت‌ها در پیکرهٔ متنی بزرگ برای دسته‌بندی حسی کلمات استفاده شد. ترنی [1] تعدادی از الگوهای زبانی برای استخراج عبارات حسی را مشخص کرد. همچنین از موتور جستجو و معیار پی‌ام‌آی[2] برای محاسبه هم‌رخدادی و ارتباط معنایی عبارات حسی مختلف با کلمات حسی کلیدی مثبت و منفی ازپیش‌تعیین‌شده (مانند کلمات "excellent" و "poor")، استفاده شد. ولیکویچ و همکاران [21] تمام چندکلمه‌ای‌های (n-grams) موجود را از میلیون‌ها صفحه وب به‌عنوان گره‌های یک گراف در نظر گرفتند. وزن یال بین این گره‌ها بر اساس فاصلهٔ کسینوسی میان بردارهای متناظر با آن عبارات محاسبه شد؛ سپس آنها با انتشار میزان بار حسی برخی عبارات حسی نخستین[3] از طریق یال‌های گراف، بار حسی سایر عبارات را محاسبه کردند.

برخی از روش‌های استخراج واژگان حسی نیز وابسته به پیکره متنی نظرات با قالب مشخص هستند. کاجی و همکارش [20] ابتدا تکرار رخداد عبارات حسی در جملات مثبت (بخش نقاط قوت سایت) و جملات منفی (بخش نقاط ضعف سایت) را شمارش کردند. پس از آن، از معیار پی‌ام‌آی برای محاسبهٔ بار حسی عبارات گزیده استفاده کردند. با ایده‌ای مشابه، نوفرستی و شمس‌فرد از یک پیکره از واژگان حسی مربوط به داروها با استفاده از کلمات کلیدی (موجودیت‌های اثرپذیر) در متن نظرات کمک گرفتند [28]. آنها از متن نظرات موجود در بخش فواید یا عوارض داروها (در منبع ورودی) برای تعیین بار حسی مثبت یا منفی استفاده کردند.

روش‌های مبتنی بر لغت‌نامه اغلب از شبکهٔ واژگان برای تعیین روابط معنایی و محاسبهٔ بار حسی کلمات استفاده می‌کنند. کمپس و همکاران [17] بر اساس روابط هم‌معنی شبکهٔ واژگان یک گراف تشکیل دادند. آنها فرض کردند که کلمات هم‌معنی دارای بار حسی یکسانی هستند؛ سپس بر اساس فاصلهٔ کوتاه‌ترین مسیر بین کلمات (گره‌های گراف) میزان بار حسی سایر کلمات را محاسبه کردند. کانایاما و همکارش [29] نیز از روابط هم‌معنایی و تضاد در شبکهٔ واژگان استفاده کردند. آنها بار حسی مجموعه کلمات دارای حس را با استفاده از یک روش بوت استرپ[4] به سایر لغات هم‌جوار انتشار دادند. تاکامورا و همکاران [22] با استفاده از روابط معنایی مختلف (از قبیل هم‌معنی، تضاد و روابط سلسله مراتبی) و هم‌رخدادی عبارات بخش توصیف[5] در شبکهٔ واژگان، شبکه‌ای معنایی از واژگان ایجاد و سپس با استفاده از یک مدل چرخشی[6] بار حسی را بین کلمات مختلف پخش کردند. رائو و همکارش [24] نیز از شبکهٔ واژگان و فرهنگ جامع کلمات[7] در OpenOffice به‌عنوان منبع برای ساخت گراف کلمات استفاده کردند؛ سپس با بهره‌گیری از یک روش شبه‌ناظر برچسب‌های حسی را در گراف انتشار دادند. در بسیاری از مقالات [23,30-32] از روش‌های مبتنی بر گام تصادفی[8] و رتبه‌دهی صفحات[9] در گراف برای تعیین بار حسی (گرایش معنایی[10]) کلمات استفاده شده‌است.

برای ایجاد شبکهٔ واژگان حسی انگلیسی به نام سنتی‌وردنت[11]، با یک الگوریتم یادگیر شبه‌ناظر در چهار مرحله به هر گروه از واژگان هم‌معنی در شبکهٔ واژگان بار حسی مثبت، منفی و میزان غیرذهنی‌بودن[12] ، انتساب دادند[33, 34]. با توجه به استفاده از این منبع برای تولید واژه‌نامهٔ حسی زبان فارسی، نحوهٔ ایجاد شبکه واژگان حسی انگلیسی در بخش‌های بعد توضیح داده‌می‌شود. نویاروسکایا و همکارش [35] برای ایجاد لغات‌نامه حسی به نام سنتی‌فول[13]، ابتدا هستهٔ نخستین از کلمات حسی از پایگاه داده افکت[14] [36] را استخراج و برای ابهام‌زدایی و آزمون درستی کلمات حسی نخستین از شبکهٔ واژگان حسی انگلیسی (سنتی‌وردنت) استفاده کردند؛ سپس با روابط معنایی مختلف موجود در شبکهٔ واژگان، مجموعهٔ لغات سنتی‌فول گسترش داده‌شد. پوریا و همکاران [25] با استفاده از یک روش دسته‌بندی شش-کلاسه برچسب‌های پایگاه دادهٔ وردنت افکت[15] [37] را به هر یک از مفاهیم پایگاه دادهٔ سنتیک‌نت[16] [38] انتساب دادند.

---

1 conjunctive

2 Pointwise Mutual Information

3 Sentiment seed words

4 boot-strapping method

5 WordNet Glosses

6 Spin Model

7 thresaurus

8 Random Walk

9 PageRank

10 Semantic orientation

11 SentiWordNet

12 objectivity

13 SentiFul

14 Affect

15 WordNet Affect

16 SenticNet

در اغلب کارهای انجام‌شده در زمینهٔ نظر کاوی در زبان فارسی، مجموعه‌ای از لغات حسی به‌صورت دستی جمع‌آوری و استفاده شده‌است [39,40]. همچنین در برخی از کارها [41,42] از ترجمهٔ فارسی (توسط مترجم ساده انگلیسی به فارسی) منابع (واژه‌نامه‌های حسی) زبان انگلیسی استفاده شده‌است. دهدار بهبهانی و همکاران [32] در روشی جدید ساخت واژگان حسی فارسی را با استفاده از منابع زبان انگلیسی (لغات حسی نخستین و شبکهٔ واژگان) پیشنهاد دادند. به این منظور آنها نخست فهرست اولیهٔ لغات حسی انگلیسی (پیکره Micro-WNOp [43]) را به‌صورت دستی مشخص کرده، سپس با استفاده از روش گام تصادفی، بار (وزن) حسی سایر لغات موجود در گراف معنایی (شبکه واژگان) را تعیین کردند. با توجه به پراکندگی[1] و ناقص‌بودن شبکهٔ واژگان فارسی موجود آنها نمی‌توانستند روش خود را به‌طورمستقیم بر روی شبکهٔ واژگان فارسی اجرا کنند. آنها سپس با استفاده از ارتباط (لینک) موجود بین گروه‌های هم‌معنی شبکهٔ واژگان انگلیسی با شبکهٔ واژگان فارس‌نت (نسخه ۱) مجموعه‌ای از واژگان حسی فارسی (UTIIS) ایجاد کردند. درنهایت، درمرحلهٔ پس‌پردازش با توجه به روابط هم‌معنایی و تضاد موجود در شبکه واژگان، بار حسی کلمات اصلاح شدند. با توجّه به تعداد کم گروه‌های هم‌معنی در فارس‌نت (نسبت به شبکه واژگان انگلیسی) سایر کلمات گروه‌های هم‌معنی شبکهٔ واژگان انگلیسی با مترجم گوگل، ترجمه و وارد مجموعهٔ لغات حسی فارسی شدند. در تمام مجموعهٔ واژگان حسی فارسی ایجادشده شامل ۱۸۱۵ کلمهٔ حسی مثبت و ۱۸۵۶ کلمهٔ حسی منفی شامل نقش‌های صفت، اسم و فعل است.

۳- شبکهٔ واژگان

شبکهٔ واژگان دانشگاه پرینستون[2] (PWN) یک پایگاه داده لغوی[3] برای زبان انگلیسی است. شبکهٔ واژگان حاوی لغات زبان طبیعی در قالب مجموعه‌های کلمات هم‌معنی[4] (گروه هم‌معنی[5]) است که در دسته‌هایی با توجه به نقش‌های نحوی‌ای مانند فعل، اسم، صفت و قید تقسیم‌بندی شده‌اند. گروه‌های هم‌معنی در هر مقولهٔ دستوری دارای روابط معنایی خاص خود هستند. به‌عنوان مثال کلمات در مقولهٔ اسم با هم دارای روابط معنایی‌ای مانند :هم‌معنایی[6] ، تضاد معنایی[7]، رابطهٔ شمول معنایی (دربرداشتن)[8]، روابط سلسله‌مراتبی[9] (جزء به کل[10] و کل به جزء[11]) و ... هستند. شبکهٔ واژگان اغلب برای ابهام‌زدایی و تعیین شباهت معنایی در کاربردهای مختلف پردازش زبان طبیعی و بازیابی اطلاعات مانند ترجمهٔ ماشینی، استخراج اطلاعات و خلاصه‌سازی و ... استفاده می‌شود. آخرین نسخهٔ پی.دبلیو.ان[12] به‌طورتقریبی شاملِ ۱۵۵۳۲۷ کلمه است که در قالب ۱۱۷۵۹۷ گروه هم‌معنی سازماندهی شدند. اخیراً در بعضی مقالات از شبکهٔ واژگان برای استخراج واژگان حسی و ویژگی‌های موجودیت موردنظر استفاده شده است [23,44].

اکنون برای بیش از چهل زبان طبیعی در جهان شبکهٔ واژگان ایجاد شده‌است که بین غالب آنها با پی.دبلیو.ان پیوند وجود دارد. در ایجاد شبکهٔ واژگان برای سایر زبان‌ها به‌طور معمول از دو روش استفاده می‌شود:

۱- ساخت شبکهٔ واژگان با استفاده از ترجمهٔ گروه‌های هم‌معنی پی.دبلیو.ان مانند شبکه واژگان فارسی دانشگاه تهران [45] و آی‌دابلیوان‌دی[13] [46] و

۲- استفادهٔ شبکهٔ واژگان از منابع زبان مقصد و روش‌های زبان‌شناسی ایجادشده و برقراری پیوند بینِ گروه‌های هم‌معنی آن با پی.دبلیو.ان مانند فارس‌نت [47] و یوروورد‌نت[14] [48].

۳-۱- شبکه‌های واژگان فارسی موجود

در زبان فارسی فعالیّت‌های زیادی در جهت ساخت شبکهٔ واژگان به‌صورت خودکار یا نیمه‌خودکار انجام شده‌است [45,47, 49-52] ولی تنها شبکهٔ واژگان فارس‌نت [47] و شبکه

[1] Sparse
[2] Princeton WordNet
[3] Electronic Lexical Database
[4] synonymous sets
[5] synset
[6] synonymy
[7] antonymy
[8] meronymy
[9] Taxonomic
[10] hyponymy
[11] hypernymy
[12] WordNet 3.1 database statistics
[13] IWND
[14] EuroWordNet

واژگان فارسی دانشگاه تهران [45] منتشر شده و برای بهره‌برداری و مقایسه قابل استفاده هستند.

۱-۱-۳- فارس‌نت

فارس‌نت (شبکهٔ واژگان عمومی زبان فارسی) یک پایگاه دانشی لغوی[1] است که حاوی اطلاعات در مورد واژه‌ها و ترکیبات زبان (مفاهیم)، اجزای کلام آنها (POS) و روابط معنایی میان آنهاست. آخرین نسخهٔ این پایگاه داده (فارس‌نت نسخه ۲,۰) برای استفاده‌های پژوهشی در دسترس است[2]. برای تولید این شبکهٔ واژگان ایده به‌کاررفته در یورووردنت استفاده شد. به این ترتیب، نخست به‌صورت دستی هسته نخست شبکهٔ واژگان فارسی ایجاد و سپس با استفاده از یک روش شبه‌ناظر در فرایندی بالا به پایین این شبکهٔ واژگان تکمیل شد. هستهٔ نخست فارس‌نت به‌کمک ترجمهٔ مفاهیم بالکانت[3] و برخی مفاهیم پرکاربرد زبان فارسی ایجاد شد. سپس با به‌کارگیری یک روش شبه‌ناظر و استفاده از منابع مختلف زبان فارسی و منابع دوزبانه (فارسی-انگلیسی) هسته نخست شبکهٔ واژگان تکمیل شد [47].

۲-۱-۳- شبکهٔ واژگان فارسی دانشگاه تهران

آخرین نسخه از شبکه واژگان فارسی تهیه‌شده در دانشگاه تهران [45] در سایت شبکه‌های واژگان چندزبانه[4] قابل دریافت است. این شبکه واژگان با اجرای الگوریتم بدون ناظر EM[5] و استفاده از یک پیکرهٔ متنی و شبکهٔ واژگان انگلیسی (PWN) ایجاد شده‌است. آنان برای محاسبهٔ احتمال نخست وجود هر کلمه را در هر گروه کلمات هم‌معنی از شبکه واژگان فارس‌نت نسخه ۱,۰ استفاده و سپس سعی کردند که این احتمال را (برای هر کلمه) به کمک روش تکراری EM بیشینه کنند.

۲-۳- فردوس‌نت

شبکهٔ واژگان ایجادشده در این پژوهش فردوس‌نت نامیده شده است. هدف اصلی از ایجاد فردوس‌نت تولید یک شبکهٔ واژگان با بازخوانی[6] بالا (پوشش تقریباً کاملی از شبکهٔ واژگان پرینستون) با دقتی مناسب برای زبان فارسی است. برای ساخت شبکهٔ واژگان فردوس‌نت، از منابع زبانی و پایگاه‌های دانش زیر استفاده شده‌است:

- شبکهٔ واژگان دانشگاه پرینستون (PWN)
- لغت‌نامه‌های دوزبانهٔ (انگلیسی-فارسی) مختلف
- مترجم متن گوگل
- پایگاه دانش ویکی‌پدیا و هستان‌شناسی یاگو[7] [53] برای برقرارکردن پیوند بین ویکی‌پدیا و پی.دبلیو.ان
- پیکره‌های متنی فارسی (چند سایت خبری و صفحات ویکی‌پدیا فارسی)
- دانش‌نامه‌ها، دایره‌المعارف و لغت‌نامه‌های فارسی
- شبکه‌های واژگان فارسی فارس‌نت و شبکه واژگان دانشگاه تهران

شبکهٔ واژگان فردوس‌نت در طی فرایند نُه‌مرحله‌ای (مطابق شکل (۱)) برای تک‌تک گروه‌های هم‌معنی، تشکیل شده‌است:

مرحله نخست: تمام کلمات گروه‌های هم‌معنی با لغت‌نامه‌های[8] دوزبانه (انگلیسی به فارسی) مختلف ترجمه می‌شوند.

مرحله دوم: به‌ازای هر گروه هم‌معنی (از شبکه واژگان انگلیسی) یک گراف دوبخشی تشکیل شده و مراحل دو الی نه انجام می‌شوند. در این گراف دوبخشی، به‌ازای هر کلمهٔ انگلیسی یک گره در سمت چپ گراف (X_i) و به‌ازای هر کلمهٔ فارسی (لیست ترجمه‌ها) یک گره در سمت راست گراف (Y_i) درنظر گرفته می‌شود؛ سپس به هر کلمهٔ انگلیسی xi و ترجمهٔ فارسی آن yj، یالی به شکل (xi,yj) بین گره مربوط رسم می‌شود. وزن هر یال وابسته به تعداد تکرار کلمات در فهرست ترجمه (با لغت‌نامه‌های مختلف) و رتبهٔ ترجمه[9] آنهاست.

مرحله سوم: این مرحله شامل دو قسمت است:

۱- نخست از پایگاه‌های دانش ویکی‌پدیا و سایر شبکه‌های واژگان فارسی، کلمات فارسی مرتبط با گروه هم‌معنی جاری استخراج می‌شود. به این منظور، نخست با استفاده از هستان‌شناسی یاگو [53] مفاهیم مربوط به کلمات گروه هم‌معنی انتخاب‌شده در ویکی‌پدیا استخراج و

---

1 lexical database

2 http://dadegan.ir/catalog/farsnet

3 BalkaNet

4 http://compling.hss.ntu.edu.sg/omw/

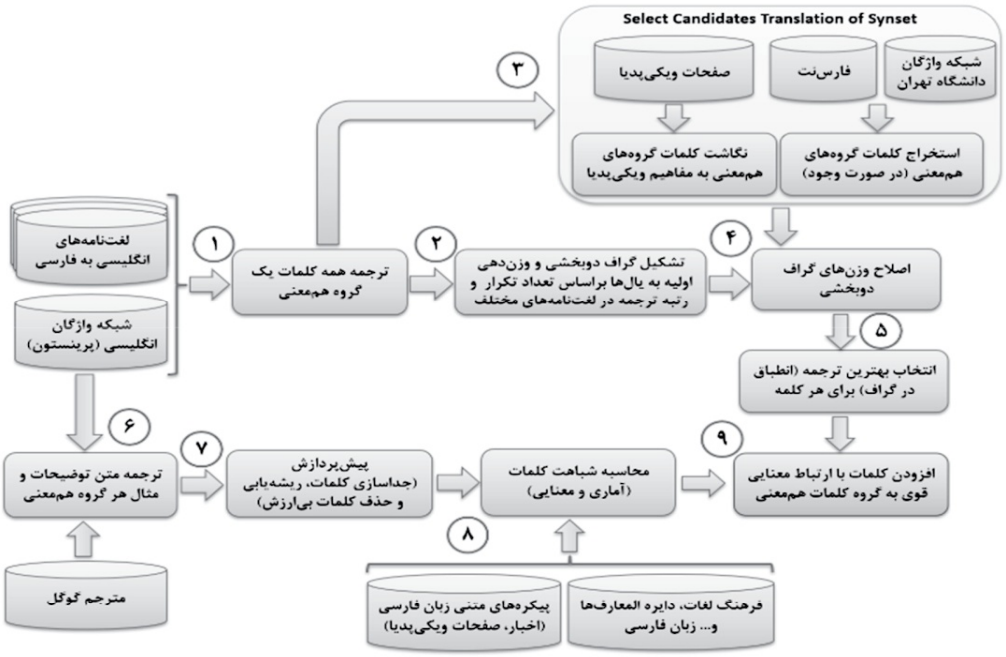
5 Expectation Maximization

6 Recall

7 Yago

8 Dictionary

۹ چون در اغلب لغت‌نامه‌های موجود کلمات ترجمه برحسب اولویت مرتب هستند.



(شکل-۱): معماری ایجاد گروه‌های هم‌معنی شبکه واژگان فردوس‌نت

(Figure-1): The system architecture for construction of the synsets of FerdowsNet

معادل فارسی این مفاهیم (در صورت وجود صفحه فارسی ویکی‌پدیا برای آن مفهوم) استخراج می‌شوند.

۲- کلمات گروه‌های هم‌معنی معادل (با استفاده از پیوند بین شبکه‌های واژگان دانشگاه تهران و فارس‌نت با پی.دبلیو.ان درصورت وجود) استخراج می‌شوند.

مرحله چهارم: در این مرحله با استفاده از کلمات استخراج‌شده در مرحلهٔ قبل (از پایگاه‌های دانش ویکی‌پدیا و سایر شبکه‌های واژگان فارسی)، کلمات ترجمهٔ گسترش (افزودن گره‌های جدید به سمت راست گراف دوبخشی)، یا وزن یال‌های مربوط به کلمات ترجمهٔ موجود در گراف دوبخشی (تشکیل شده در مرحله دوم) تقویت (افزوده) می‌شوند.

مرحله پنجم: با استفاده از الگوریتم هانگاریـن[1] (مجاری)[2] بهترین تطابق[3] (مناسب‌ترین ترجمه‌های فارسی برای کلمات انگلیسی) از گراف دو بخشی وزن‌دار تشکیل‌شده استخراج می‌شوند. به این ترتیب فهرست کلمات نامزد S برای گروه هم‌معنی فعلی به دست می‌آید. همچنین در این مرحله، از بین کلمات فارسی انتخاب‌نشده، کلماتی که وزن ارتباطی آنها (با یکی از کلمات انگلیسی) بیشتر از متوسط وزن ارتباطی کلمات انتخاب‌شده (یال‌های انتخاب‌شده) باشد، به‌عنوان ترجمهٔ گزیدهٔ دوم (هم مفهوم با کلمه انگلیسی)، انتخاب می‌شود.

مرحله ششم: توضیحات[4] و مثال‌های مربوط به هر گروه هم‌معنی (در پی.دابلیو.ان) با مترجم گوگل ترجمه می‌شود.

مرحله هفتم: پیش‌پردازش لازم بر روی متن ترجمه انجام‌شده و کلمات کلیدی متن (معادل با کلمات فارسی در سمت راست گراف دوبخشی) استخراج می‌شوند.

مرحله هشتم: با استفاده از پیکره‌های متنی زبان فارسی (خبرگزاری‌های برخط مختلف [56-54] و محتویات صفحات ویکی‌پدیا فارسی) و همچنین فرهنگ واژگان مختلف فارسی، کلمات هم‌معنی و هم‌کاربرد (با استفاده از معیار پی‌ام‌آی [1,57]) با کلمات انتخاب‌شده در مرحله پنجم (مجموعه کلمات S) استخراج می‌شوند.

مرحله نهم: از میان کلمات استخراج‌شده در مرحلهٔ پیش، کلماتی که شباهت آنها با کلمات مجموعه S (که در مرحله پنجم انتخاب شدند) بیش از حد آستانهٔ مشخصی باشد، به فهرست کلمات نهایی افزوده می‌شوند.

پس از ترجمه و گسترش کلمات گروه‌های هم‌معنی

[1] Hungarian https://github.com/KevinStern/software-and-algorithm

[2] s/blob/master/src/main/java/blogspot/software_and_algorithms/stern_library/optimization/HungarianAlgorithm.java

[3] Matching

[4] Glossary

پی.دابلیو.ان، از روابط موجود بین گروه‌های هم‌معنی در این شبکهٔ واژگان برای ارتباط بین گروه‌های هم‌معنی فردوس‌نت استفاده می‌شود.

در شبکهٔ واژگان فردوس‌نت برای هر کلمه در گروه کلمات هم‌معنی مقداری به‌عنوان میزان اطمینان[1] با نگاشت وزن یال‌های مرتبط با آن کلمه در گراف دوبخشی کلمات (که در الگوریتم ساخت فردوس‌نت توضیح داده‌شد)، به عددی بین صفر و یک، محاسبه می‌شود.

۴- تولید شبکهٔ واژگان حسی

بدیهی است که تشخیص عبارات حسی و تعیین میزان حس آنها (کمی‌سازی میزان حس) بدون کمک انسان، برای ماشین غیرممکن است. بنابراین در روش‌های تحلیل حس، ابتدا فهرستی از عبارات حسی نخستین که به‌طور معمول دارای مقدار عددی تعیین‌کننده در میزان بار حسی هستند، با کمک اشخاص خبره تهیه شده و به‌عنوان ورودی به تحلیل‌گر حس متن داده‌می‌شود؛ سپس با الگوریتم‌های مختلفی فهرست نخستین عبارات حاوی حس بسط و تکمیل می‌شود و میزان و شدت حس نیز با توجه به برخی کلمات جمله از قبیل منفی‌کننده‌ها (معکوس‌کننده‌های حسی) تنظیم می‌شود. افعال اسنادی منفی (از قبیل نیست، نبود و نمی‌باشد) و برخی کلمات (مانند «به‌ندرت»، «عدم وجود» و «بدون») از مهم‌ترین معکوس‌کننده‌های حسی جملات در زبان فارسی هستند. همان‌طورکه اشاره شد، تولید واژه‌نامهٔ لغات حسی یکی از بخش‌های اساسی و مهم برای تشخیص حس و شدت آن است. برای ایجاد مجموعهٔ واژگان حسی به طور عموم دو روش در مقالات استفاده شده‌است:

۱- توسعه یا ترجمهٔ عبارات حسی از روی مجموعهٔ واژگان حسی موجود [60-58]

۲- گسترش فهرست کلمات حسی نخستین با استفاده از شبکهٔ واژگان، پیکره‌ای از متون نظرات و سایر منابع زبانی [33,61-63].

اغلب در رویکرد دوم، فهرست کلمات حسی نخستین به‌صورت دستی تهیه می‌شوند. در این مقاله با ایجاد یک شبکهٔ واژگان کامل برای زبان فارسی و برقراری پیوند بین مفاهیم آن با شبکهٔ واژگان از رویکرد نخست استفاده شده‌است. به عبارت دیگر، برای تولید مجموعهٔ واژگان حسی فارسی از بار حسی کلمات در شبکهٔ واژگان حسی انگلیسی (SentiWordNet) استفاده شده‌است.

روش پیشنهادی این مقاله بر پایهٔ استفاده از بار حسی کلمات موجود در واژه‌نامهٔ شبکهٔ واژگان حسی انگلیسی نسخه ۳,۰ با نگاشت و ارتباط بین فردوس‌نت و شبکهٔ واژگان انگلیسی است.

۴-۱- ابزارهای پردازش متن فارسی

زبان فارسی یکی از زبان‌های هندو-اروپایی[2] است که بیش از صد میلیون نفر از مردم جهان با آن صحبت می‌کنند و به‌عنوان زبان رسمی سه کشور ایران، افغانستان (فارسی دری) و تاجیکستان (فارسی تاجیکی) است. به‌دلیل پیچیدگی‌های زبانی، منابع و مطالعات انجام‌شده در این زبان از دیدگاه محاسباتی پژوهش‌گران کمتر به آن توجه کرده‌اند [64,65]. استانداردسازی، تفکیک متن به جملات، عبارات و کلمات و برچسب‌گذاری و حاشیه‌نویسی آنها تأثیر به‌سزایی بر روی پردازش و استخراج اطلاعات، دسته‌بندی یا دیگر کاربردهای پردازش زبان طبیعی دارد. با توجه به اهمیت مرحلهٔ پیش‌پردازش در تحلیل نظرات [66]؛ مجموعه‌ای از ابزارهای کاربردی پردازش متون زبان فارسی برای استفاده در شبکهٔ واژگان حسی فارسی تهیه شده‌است.

۴-۲- شبکهٔ واژگان حسی انگلیسی

شبکهٔ واژگان حسی انگلیسی [33] یکی از بهترین منابع موجود برای شناسایی کلمات حسی است که بر اساس تعیین میزان بار حسی هر کدام از گروه هم‌معنی در شبکهٔ واژگان انگلیسی پرینستون (PWN) ایجاد شده‌است. شبکهٔ واژگان حسی انگلیسی برای هر کدام از گروه‌های هم‌معنی میزان بار حسی منفی[3]، مثبت[4] و همچنین مقدار غیرحسی[5] بودن (با توجه به مقدار حس مثبت و منفی)[6] را با عددی بین صفر و یک مشخص می‌کند.

شبکهٔ واژگان حسی انگلیسی نسخه ۱/۰[7] [34] با یک الگوریتم یادگیر شبه‌ناظر در چهار مرحله تشکیل شده‌است:

۱- بار حسی مثبت و منفی تعدادی محدودی از گروه‌های

[1] Confidence

[2] Indo-European

[3] negativity

[4] positivity

[5] objectivity

[6] 1-(positivity score + negativity score)

[7] SentiWordNet v1.0

هم‌معنی اولیه به‌صورت دستی مشخص شده و این بار حسی با توجه به روابط معنایی (هم‌معنی و متضاد) موجود در شبکهٔ واژگان بر روی کلمات و گروه‌های هم‌معنی مرتبط بسط داده‌می‌شود.

۲- سپس تعدادی از گروه‌های هم‌معنی فاقد بار حسی[1] نیز مشخص می‌شوند (به‌صورت دستی برچسب‌گذاری می‌شوند) و توضیحات[2] آنها به‌همراه گروه‌های هم‌معنی مشخص‌شده در مرحلهٔ پیش به‌عنوان داده‌های آموزشی برای استفاده در مرحلهٔ یادگیری یک روش دسته‌بندی باناظر به‌کارگرفته‌می‌شوند.

۳- با الگوریتم دسته‌بندی (با نرخ بازخوانی پایین و دقت بالا) سایر گروه‌های هم‌معنی برچسب‌گذاری حسی[3] می‌شوند.

۴- به‌منظور کاستن از خطای الگوریتم‌های دسته‌بندی، گروه‌های هم‌معنی آماده‌شده در مرحله دو برای آموزش چند الگوریتم دسته‌بندی استفاده می‌شوند. بعد از اجرای آنها (مرحله سوم)، در این مرحله نتایج آنها با یکدیگر ترکیب می‌شوند. بدین منظور به‌ازای هر گروه هم‌معنی کلمات، میانگین بار حسی به‌دست‌آمده از روش‌های دسته‌بندی مختلف محاسبه شده و به عددی در بازه صفر الی یک نگاشت می‌شود.

سپس در نسخهٔ سوم این شبکه واژگان حسی [33] با استفاده از الگوریتم تکراری گام تصادفی بر روی گراف شبکه واژگان پی.دابلیو.ان، نتایج حاصل از نسخهٔ پیش (SentiWordNet v1.0) را اصلاح کردند.

تاکنون شبکهٔ واژگان حسی انگلیسی به‌عنوان یک منبع واژگان حسی مستقل از دامنه و موضوع در بسیاری از کاربردهای نظرکاوی استفاده شده‌است [44, 67,68]. علاوه بر این، با توجّه به رابطهٔ این واژه‌نامه با پی.دابلیو.ان، از این منبع در بسیاری از کاربردهای نظرکاوی در زبان‌های دیگر نیز (با برقرار کردن پیوند بین شبکه‌های واژگان آن زبان‌ها با پی.دابلیو.ان) استفاده شده‌است [58,61].

۳-۴- ساخت واژه‌نامهٔ حسی زبان فارسی

همان‌طورکه اشاره شد، در این پژوهش برای تولید حس‌نگار ابتدا با استفاده از نگاشت مفاهیم (گروه‌های هم‌معنی) در شبکهٔ واژگان پرینستون به زبان فارسی، شبکهٔ واژگان جامع (فردوس‌نت) ساخته شده‌است. درنهایت، با استفاده از فردوس‌نت، میزان بار حسی محاسبه‌شده برای هر گروه هم‌معنی در شبکهٔ واژگان حسی انگلیسی به گروه‌های هم‌معنی متناظر با آن در حس‌نگار نگاشت می‌شود. پس درواقع با ابهام‌زدایی مفاهیم شبکهٔ واژگان حسی انگلیسی، یک شبکهٔ واژگان حسی برای زبان فارسی ایجاد شده‌است.

به این ترتیب واژه‌های حسی زبان فارسی در چهار مقولهٔ صفت (مانند توانا، زشت)، اسم (مانند پیش‌کسوت، درد)، قید (مانند عجولانه، به‌آرامی) و فعل (مانند عشق ورزیدن، خندیدن) دسته‌بندی می‌شوند.

از شبکهٔ واژگان حسی فارسی به‌عنوان یک واژه‌نامه حسی مرجع برای زبان فارسی می‌توان استفاده کرد. واژه‌نامهٔ حسی فارسی به‌دست‌آمده از کلمات حسی با میزان بار حسی بیشتر از ۰/۵ (بدون در نظر گرفتن میزان اطمینان[4])، شامل ۶۰۶۳ واژه حسی مثبت و ۹۴۴۱ واژه حسی منفی است.

علاوه‌براین، با توجّه به وجود درجهٔ اطمینان برای کلمات موجود در هر گروه کلمات هم‌معنی در فردوس‌نت، برای هر کلمهٔ حسی علاوه بر بار حسی مثبت و منفی، میزان اطمینان (اعتبار) نیز خواهیم‌داشت. تعداد کلمات حسی با نقش‌ها و میزان اطمینان مختلف در جدول (۱) گردآوری شده‌است. تأثیر میزان اطمینان بر دقت و بازخوانی شبکهٔ واژگان فردوس‌نت و درنتیجه تأثیر آن بر واژه‌نامهٔ حسی در بخش ارزیابی نشان داده شده‌است.

۵- ارزیابی

برای ارزیابی واژگان حسی استخراج‌شده، مجموعهٔ نظرات وبگاه دیجی‌کالا[5] با خزش‌گر وب[6] در سال ۱۳۹۲ جمع‌آوری شده‌است. دیجی‌کالا پنجمین سایت پربازدید ایران و بزرگترین فروشگاه اینترنتی[7] در ایران و خاورمیانه است[8] و با توجه به حجم بالای کاربران، از غنای نظرات تقریباً خوبی

---

[1] objective

[2] The glosses of the synsets

[3] برچسب‌های "pos"، "neg" و "obj"

[4] در صورتی که مجموع بار حسی مثبت و منفی کلمه بیشتر از ۰،۵ باشد، در این صورت کلمه حسی و در غیر این صورت وزن غیرحسی کلمه بیشتر است.

[5] www.digikala.com

[6] Web crawler

[7] market leader in e-commerce

[8] http://www.alexa.com/topsites/countries/IR

(جدول-۱): تعداد کلمات حسی مثبت (Pos#) و منفی (Neg#) موجود در شبکهٔ واژگان حسی فارسی (حس‌نگار) براساس محدودیت‌های مختلف

(Table-1): Positive (Pos#) and negative (Neg#) sentiment words in PSWN

		اسم	صفت	قید	فعل	همه
بدون محدودیت (همه)	Pos#	۱۱۱۹۸	۹۴۹۱	۲۳۳۴	۳۶۷۵	۲۶۶۹۸
	Neg#	۱۳۱۷۵	۱۱۰۲۲	۶۹۸	۴۱۷۳	۲۹۰۶۸
میزان اطمینان بالای ۰٫۸	Pos#	۱۷۴۹	۱۳۷۲	۴۰۸	۳۸۳	۳۹۱۲
	Neg#	۲۱۰۰	۱۶۲۴	۱۱۹	۴۷۱	۴۳۱۴
میزان اطمینان بالای ۰٫۸ و بار حسی بیشتر از ۰٫۵	Pos#	۳۶۶	۴۸۳	۲۴	۳۹	۹۱۲
	Neg#	۵۵۳	۷۵۰	۱۵	۸۷	۱۴۰۵

برخوردار است. این مجموعهٔ داده شامل نظرات متنی بیان‌شده درباره محصولات مختلف است.

تعداد کل نظرات پیکره جمع‌آوری‌شده شامل ۳۱۷۳۰ نظر در ده نوع کالای متفاوت است. داوران حدود ۳۰۸۰ متن نظرات را برای آموزش الگوریتم‌های یادگیری ماشین باناظر و ارزیابی روش‌های نظرکاوی، برچسب زده‌اند؛ ولی بقیه نظرات فاقد برچسب حسی هستند. برخی از ویژگی‌های این مجموعه داده در جدول (۲) فهرست شده‌اند.

این مجموعه داده دارای سه دسته نظر مختلف مربوط به «نقد خبره»، «نظرات کاربران فعال» و «نظرات کاربران عادی و مهمان» است. در جدول (۳) ویژگی هر دسته از نظرات و تفاوت‌های آنها قابل‌مشاهده است.

(جدول-۲): ویژگی‌های پیکره نظرات دیجیکالا

(Table-2): Features of Digikala review corpus

	کل پیکره	دارای برچسب حسی
تعداد انواع (گروه) محصولات	۱۰	۱۰
تعداد محصولات (کالاها)	۷/۵۷۲	۲۷۰
تعداد نظرات	۳۱/۷۳۰	۳/۰۸۰

(جدول-۳): ویژگی‌های انواع نظرات در پیکره دیجی‌کالا

(Table-3): A variety of opinions available on the corpus

	نحوه بیان (نوع نوشتار)	ویژگی‌های مورد بررسی	میزان ذهنی[1] و سلیقه‌ای بودن	تعداد متوسط نظر به ازای هر کالا	طول (تعداد کاراکتر) متوسط متن
مسئول سایت	رسمی	انواع ویژگی‌ها	نسبتا کم[2]	۱	۸۰۴۵
کاربران فعال	نیمه رسمی	تعداد زیاد	متوسط	۳/۷۹	۱۷۴۳
کاربران عادی	عامیانه	تعداد محدود	زیاد	۹/۱۲	۳۵۶

۱-۵- ارزیابی شبکهٔ واژگان فردوس‌نت و حس‌نگار

ابتدا شبکهٔ واژگان فردوس‌نت به‌صورت کمّی ارزیابی و با سایر شبکه‌های واژگان موجود در زبان فارسی مقایسه می‌شود. در جدول (۴) ویژگی‌های شبکه‌های واژگان فارسی مختلف گردآوری شده‌است.

برای ارزیابی کیفی فردوس‌نت و مقایسه آن با دیگر شبکه‌های واژگان موجود در زبان فارسی، ابتدا حدود هزار گروه هم‌معنی از شبکهٔ واژگان انگلیسی (از هریک از مقوله‌های اسمی، فعل، صفت و قید، ۲۵۰ گروه هم‌معنی) به‌صورت تصادفی انتخاب و سپس کلمات گروه‌های هم‌معنی معادل آنها در شبکه‌های واژگان مختلف زبان فارسی استخراج و برای داوری به اشخاص خبره داده‌شدند؛ سپس ارزیابی کیفی کلمات درون گروه‌های هم‌معنی مختلف توسط داوران خبره در زمینهٔ پردازش زبان طبیعی انجام شده‌است. برای ارزیابی دقیق کیفیت، معیارهای دقت و بازخوانی برای هر شبکه واژگان به‌صورت جداگانه محاسبه شده‌است.

بدین منظور ابتدا به کمک داوران، به‌ازای هر گروه کلمات هم‌معنی انگلیسی یک مجموعه مرجع از لغات فارسی معادل آن (مفهوم) به‌صورت $S^*=\{s_1^*, s_2^*, s_3^*, \dots, s_n^*\}$ در نظر گرفته‌می‌شود. همچنین، مجموعه واژه‌های موجود در شبکه واژگان برای هر گروه هم‌معنی، $S^{wn}=\{s_1^{wn}, s_2^{wn}, s_3^{wn}, \dots,$

[1] subjective

[۲] اغلب بیان ویژگی‌های کمّی (objective)

(جدول-۴): ارزیابی کمّی شبکه‌های واژگان زبان فارسی
(Table-4): Quantitative assessment Persian WordNets

	شبکه واژگان انگلیسی	فارس‌نت ۲	شبکه واژگان فارسی (دانشگاه تهران)	فردوس‌نت
تعداد کلمات (یکتا)	۱۵۵۲۸۷	۳۱۲۳۰	۱۸۱۶۶	۱۰۰۰۶۲
تعداد گروه‌های هم‌معنی	۱۱۷۶۵۹	۲۰۴۳۲	۱۷۷۵۹	۹۱۶۴۰
تعداد گروه کلمات هم‌معنی لینک شده با شبکه واژگان انگلیسی	۱۱۷۶۵۹	۱۷۳۰۰	۱۷۷۵۹	۹۱۶۴۰
متوسط تعداد کلمه در هر گروه کلمات هم‌معنی	۱/۷۵۹	۱/۸۵۳	۱/۷۱۵	۲/۰۳۳
درصد گروه‌های هم‌معنی تک‌کلمه‌ای	۵۴٪	۵۶٪	۵۸٪	۶۰٪
تعداد متوسط اشتراک هر کلمه در گروه‌های هم‌معنی مختلف	۱/۵۱۴	۱/۱۶	۱/۵۷	۱/۸۸

$\{s_n^{wn}$ نامیده می‌شود. درنهایت دقت و بازخوانی به‌ازای هر یک از گروه‌های کلمات هم‌معنی به‌صورت زیر محاسبه شده و میانگین آنها به‌عنوان دقت و بازخوانی نهایی شبکه واژگان در نظر گرفته می‌شود.

رابطه ۱
$$Recall = \frac{|S^* \cap S^{wn}|}{|S^*|}$$

رابطه ۲
$$Precision = \frac{|S^* \cap S^{wn}|}{|S^{wn}|}$$

رابطه ۳
$$F_1 - Measure = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

عبارت $|S^* \cap S^{wn}|$ بیان‌کنندهٔ تعداد کلمات درست انتخاب‌شده در گروه کلمات هم‌معنی موردنظر است. عبارت $|S^*|$ تعداد کلمات فارسی معادل با مفهوم اصلی و $|S^{wn}|$ تعداد کلمات در گروه هم‌معنی مورد نظر در شبکه واژگان فارسی است.

در مرحلهٔ بعد، کلمات و توضیحات هر گروه هم‌معنی در شبکه واژگان انگلیسی در کنار فهرست لغات فارسی (معادل با آن گروه هم‌معنی) به شکل یک فهرست آماده شد. فهرست لغات فارسی به‌طوری بود که برای داوران مشخص نبود که هر کلمهٔ فارسی متعلق به کدام شبکهٔ واژگان است. درنهایت تعدادی از داوران خبره (در زمینهٔ مفاهیم پردازش زبان طبیعی و شبکه واژگان) کلمات اشتباه هر گروه هم‌معنی را مشخص کردند. کلمات اشتباه کلماتی هستند که معادل مفهوم بیان شده در گروه هم‌معنی انگلیسی نباشند. سپس با استفاده از برچسب‌های داوران میزان دقت و بازخوانی و F1-Measure برای هر یک از گروه‌های هم‌معنی محاسبه شده‌است. نتایج ارزیابی دقت، بازخوانی و مقدار F1-Measure شبکه‌های واژگان مختلف برای مجموعه هزار عددی از گروه‌های هم‌معنی در جدول (۵) نمایش داده شده‌است.

با توجه به وجود درجه اطمینان به‌ازای هر یک از کلمات موجود در گروه‌های هم‌معنی در شبکهٔ واژگان فردوس‌نت، در این شبکهٔ واژگان ارزیابی کیفی به‌ازای بازه‌های اطمینان مختلف محاسبه شده‌است. در جدول (۵)، منظور از *conf* درجهٔ اطمینان کلمات درون هر گروه هم‌معنی در شبکهٔ واژگان فردوس‌نت است. همچنین باید توجه شود که در شبکه‌های واژگان فارسی، گروه‌های هم‌معنی فارسی معادل با برخی از گروه‌های هم‌معنی موجود در شبکهٔ واژگان انگلیسی وجود ندارد. به این منظور در جدول (۶) یک‌بار بازخوانی برای این گروه‌ها صفر و یک‌بار نیز این گروه‌ها در محاسبه میانگین بازخوانی (موجود) کلی در نظر گرفته نشده‌اند. به‌عنوان مثال، برای شبکهٔ واژگان فارس‌نت، در مجموعه هزار عددی گروه‌های هم‌معنی انتخاب‌شده از شبکهٔ واژگان انگلیسی برای ارزیابی، تنها تعداد ۱۴۷ (حدود ۱۵٪) گروه هم‌معنی معادل در شبکهٔ واژگان فارس‌نت وجود داشتند. دقت کلمات درون گروه‌های هم‌معنی موجود حدود ۰/۸۹۸ و میزان بازخوانی کلی (از هزار گروه هم‌معنی) ۰,۱۰۱ است؛ ولی اگر مقدار بازخوانی تنها برای همان ۱۴۷ گروه هم‌معنی (گروه‌های هم‌معنی که معادل آنها در فارس‌نت موجود است) محاسبه شود، مقدار بازخوانی این شبکهٔ واژگان حدود ۰/۶۹۵ خواهدبود.

برای ارزیابی کیفی شبکهٔ واژگان حسی فارسی (حس‌نگار) و واژه‌نامهٔ حسی UTIIS، داوران حدود ۱۵۰ کلمهٔ حسی را از متن نظرات استخراج کردند. با رویکردی مشابه پیش، دقت و بازخوانی شبکهٔ واژگان حسی، محاسبه و نتایج آن در جدول (۶) بیان شده‌است.

برای محاسبهٔ بازخوانی مشخص شد که از بین ۱۵۰ عبارت حسی، ۱۱۳ مورد در بین لغات شبکه واژگان حسی (حس‌نگار) وجود داشتند. پس به عبارت دیگر، میزان بازخوانی شبکهٔ واژگان حسی حدود ۰/۷۵ است. البته اغلب عبارات

(جدول-۵): ارزیابی کیفی شبکه‌های واژگان زبان فارسی
(Table-5): The qualitative assessment of Persian WordNets

شبکه‌های واژگان زبان فارسی	پوشش گروه‌های هم‌معنی شبکه واژگان انگلیسی	دقت	بازخوانی	F1-Measure	بازخوانی موجود	F1- Measure موجود
فارس‌نت ۲	۱۵٪	۰/۸۹۸	۰/۱۰۱	۰/۱۸۲	۰/۶۹۵	۰/۷۸۴
شبکه واژگان فارسی دانشگاه تهران	۱۱٪	۰/۷۹۷	۰/۰۵۷	۰/۱۰۶	۰/۵۳۱	۰/۶۳۷
فردوس‌نت ($conf \geq 0.8$)	۴۵٪	۰/۸۲۶	۰/۱۹۹	۰/۳۲۱	۰/۴۸۳	۰/۶۱
فردوس‌نت ($conf \geq 0.4$)	۵۶٪	۰/۸۰۱	۰/۳۱۰	۰/۴۴۷	۰/۶۲۷	۰/۷۰۳
فردوس‌نت (همه کلمات)	۸۲٪	/۷۴۰	۰/۶۰۳	۰/۶۶۵	۰/۹۰۱	۰/۸۱۳

حسی که در شبکه واژگان وجود ندارند، مربوط به کلمات عامیانه یا اشکلات املایی کلمات حسی متن نظراتی هستند که توسط ابزارهای پیش‌پردازش متن تولیدشده، تصحیح نشده‌اند. از میان ۱۱۳ کلمه موجود در شبکه واژگان حسی، نوع حس (از نظر مثبت یا منفی بودن)، ۹۷ کلمه مطابق با حس بیان‌شده توسط داوران بود. پس دقت شبکه واژگان حسی فارسی ایجادشده، حدود ۰/۸۶ است. برای کلماتی که در چند گروه هم‌معنی حضور داشتند، متوسط بار حسی آنها در گروه‌های مختلف درنظر گرفته شده‌است.

(جدول-۶): ارزیابی کیفی واژه‌نامه حسی زبان فارسی
(Table-6): Qualitative assessment of the Persian sentiment lexicons

	دقت	بازخوانی	F1-Measure
شبکه واژگان حسی فارسی (حس‌نگار)	۰/۸۶	۰/۷۵	۰/۸
واژه‌نامه حسی UTIIS [32]	۰/۸۷۷	۰/۶۵۳	۰/۷۵

بخشی از خطای شبکه واژگان حسی فارسی مربوط به خطای موجود در شبکه واژگان فارسی فردوس‌نت است. علاوه‌بر آن بار حسی مشخص‌شده برای کلمات در شبکه واژگان حسی انگلیسی نیز دارای مقداری خطا است که این خطا در شبکهٔ واژگان حسی فارسی نیز انتشار می‌یابد و دقت آن را کاهش می‌دهد [69].

۶- نتیجه‌گیری

در این پژوهش ابزارهای پردازش متن، شبکهٔ واژگان فردوس‌نت و به دنبال آن شبکه واژگان حسی زبان فارسی ایجاد و معرفی شدند. برای تولید شبکه واژگان حسی (واژه‌نامه حسی) فارسی از روش نگاشت و استفاده از شبکه واژگان حسی انگلیسی برای زبان فارسی، با استفاده از طراحی شبکه واژگان فردوس‌نت (و سایر منابع زبانی) پیشنهاد شد. علاوه‌بر این، برای افزایش کیفیت روش‌های تحلیل نظرات، مجموعه‌ای از ابزارهای پیش‌پردازش زبان فارسی ایجاد و توسعه داده‌شدند. همچنین در این راستا، پیکرهٔ نظرات محصولات تجاری برای نظرکاوی در زبان فارسی توسعه پیدا کرد[1]؛ سپس ارزیابی دقیقی بر روی کیفیت ابزارهای پردازش متن و شبکهٔ واژگان فردوس‌نت (در مقایسه با سایر شبکه‌های واژگان موجود برای زبان فارسی) انجام شد. درنهایت نرخ دقت و بازخوانی شبکه واژگان حسی فارسی نیز تخمین زده‌شد.

این مقاله علاوه‌بر معرفی ابزارهای پردازش زبان فارسی، مجموعه واژگان و الگوهای حسی مناسب و پیکره نظرکاوی فارسی، مبنای خوبی برای انتخاب و استفاده از ویژگی‌ها، روش‌های انتخاب ویژگی و روش‌های مختلف دسته‌بندی برای ادامه پژوهش‌های نظرکاوی در زبان فارسی (و سایر زبان‌های مشابه مانند اردو و عربی) برای پژوهش‌گران این حوزه می‌تواند باشد.

سپاس‌گزاری

در این قسمت، لازم است از زحمات دانشجویان گروه زبان‌شناسی و اعضای آزمایشگاه فناوری تارنمای دانشگاه فردوسی مشهد که در برچسب‌زنی پیکره نظرات و ارزیابی ابزارهای پردازش متون مختلف مشارکت داشتند، سپاس‌گزاری شود.

۷- مراجع
7-References

[1] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for*

[1] برای استفاده‌های پژوهشی (غیرتجاری) از پیکره و ابزارهای پردازش متن فارسی می‌توانید به تارنمای آزمایشگاه فناوری وب دانشگاه فردوسی (http://wtlab.um.ac.ir) مراجعه نمایید.

computational linguistics, 2002, pp. 417-424.

[2] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005.

[3] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," presented at the Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003.

[4] S.-M. Kim and E. Hovy, "Extracting opinions, opinion holders, and topics expressed in online news media text," presented at the Proceedings of the Workshop on Sentiment and Subjectivity in Text, 2006.

[5] C. O. Alm, "Subjective natural language problems: motivations, applications, characterizations, and implications," presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, 2011.

[6] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010.

[7] M. Abdul-Mageed, M. Diab, and M. Korayem, "Subjectivity and sentiment analysis of modern standard Arabic," presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.

[8] I. Habernal, T. Ptáček, and J. Steinberger, "Supervised sentiment analysis in Czech social media," *Information Processing & Management,* vol. 50, pp. 693-707, 2014.

[9] H. Guo, H. Zhu, Z. Guo, X. Zhang, and Z. Su, "OpinionIt: a text mining system for cross-lingual opinion analysis," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 1199-1208.

[10] D. Gao, F. Wei, W. Li, X. Liu, and M. Zhou, "Cross-lingual Sentiment Lexicon Learning With Bilingual Word Graph Label Propagation," *Computational Linguistics,* vol. 41, pp. 21-40, 2015.

[11] M.-T. Martín-Valdivia, E. Martínez-Cámara, J.-M. Perea-Ortega, and L. Alfonso Ureña-López, "Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches," *Expert Systems with Applications,* vol. 40, pp. 3934-3942, 2012.

[12] C. Banea, R. Mihalcea, and J. Wiebe, "Porting Multilingual Subjectivity Resources Across Languages," *IEEE Transactions on Affective Computing,* vol. 4, 2013.

[13] A. Balahur and M. Turchi, "Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis," *Computer Speech & Language,* vol. 2 ,۸pp. 56–75, 2013.

[14] M. Okada and K. Hashimoto, "Investigation of Preprocessing of Multilingual Online Reviews for Automatic Classification," in *Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on*, 2012, pp. 306-30.۹

[15] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the international conference on Web search and web data mining*, 2008, pp. 231-240.

[16] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics,* vol. 37, pp. 267-307, 2011.

[17] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 2004, pp. 1115-1118.

[18] A. Fahrni and M. Klenner, "Old wine or warm beer: Target-specific sentiment analysis of adjectives," in *Proc. of the Symposium on Affective Language in Human and Machine, AISB*, 2008, pp. 60-63.

[19] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the eighth confer-ence on European chapter of the Associa-tion for Computational Linguistics*, 1997, pp. .۱۸۱-۱۷٤

[20] N. Kaji and M. Kitsuregawa, "Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents," in *EMNLP-CoNLL*, 2007, pp. 1075-1083.

[21] L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. McDonald, "The viability of web-derived polarity lexicons," in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 777-785.

[22] H. Takamura, T. Inui, and M. Okumura, "Extracting semantic orientations of words using spin model," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 133-140.

[23] A. Esuli and F. Sebastiani, "Pageranking wordnet synsets: An application to opinion mining," presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL), Prague, Czech Republic, 2007.

[24] D. Rao and D. Ravichandran, "Semi-supervised polarity lexicon induction," in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 675-682.

[25] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, and S. Bandyopadhyay, "Enhanced SenticNet with affective labels for concept-based opinion mining," *IEEE Intelligent Systems,* vol. 28, pp. 31-38, 2013.

[26] S. Gindl, A. Weichselbraun, and A. Scharl, "Extracting and Grounding Contextualized Sentiment Lexicons," 2013.

[27] D. Tang, F. Wei, B. Qin, M. Zhou, and T. Liu, "Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach," in *the 25th International Conference on Computational Linguistics (COLING)*, 2014, pp. 172-182.

[۲۸] س. نوفرستی، س.، و م. شمس فرد، "ساخت نیمه‌خودکار یک پیکره از نظرات غیرمستقیم در دامنه دارو و به‌کارگیری آن برای تعیین قطبیت نظرات". چاپ شده در مجله پردازش و علائم داده‌ها، دوره ۱۳ (۲)، سال ۱۳۹۵، ۳۵-۴۹.

[28] S. Nofersti and M. Shamsfard, "Automatic building a corpus and exploiting it for polarity classification of indirect opinions about drugs.", in *Journal of Signal and Data Processing (JSDP)*, 2016; 13 (2), pp.35-49.

[29] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006, pp. 355-363.

[30] A. Hassan and D. Radev, "Identifying text polarity using random walks," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 395-403.

[31] A. Hassan, A. Abu-Jbara, R. Jha, and D. Radev, "Identifying the semantic orientation of foreign words," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, 2011, pp. 592-597.

[32] I. Dehdarbehbahani, A. Shakery, and H. Faili, "Semi-supervised word polarity identification in resource-lean languages," *Neural Networks,* vol. 58, pp. 50-59, 2014.

[33] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in *LREC*, 2010, pp. 2200-2204.

[34] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, 2006, pp. 417-422.

[35] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "SentiFul: A lexicon for sentiment analysis," *IEEE Transactions on Affective Computing,* vol. 2, pp. 22-36, 2011.

[36] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Textual affect sensing for sociable and expressive online communication," in *International Conference on Affective Computing and Intelligent Interaction*, 2007, pp. 218-229.

[37] C. Strapparava and A. Valitutti, "WordNet Affect: an Affective Extension of WordNet," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, 2004, pp. 1083-1086.

[38][E. Cambria, R. Speer, C. Havasi, and A. Hussain, "SenticNet: A Publicly Available Semantic Resource for Opinion Mining," in *AAAI fall symposium: commonsense knowledge*, 2010.

[39] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghassem-Aghaee, "A Framework for Sentiment Analysis in Persian," *Open Transactions on Information Processing,* vol. 1, pp. 1-14, 2014.

[40] F. Amiri, S. Scerri, and M. H. Khodashahi, "Lexicon-based Sentiment Analysis for Persian Text," in *Recent Advances in Natural Language Processing*, 2015, pp. 9-16.

[41] M. Shams, A. Shakery, and H. Faili, "A non-parametric LDA-based induction method for sentiment analysis," in *Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on*, 2012, pp. 216-221.

[۴۲] س.ع. مردانی و ع. آقایی، "ارائهٔ روش نظارتی برای نظرکاوی در زبان فارسی با استفاده از لغت‌نامه و الگوریتم SVM"، فصلنامه علمی-پژوهشی مدیریت فناوری اطلاعات, دوره ۷ (۲)، ۱۳۹۳، ۳۶۲-۳۴۵.

[42] A.Mardani and S.A.Aghaie "A superviesd method for opinion mining in Persian using lexicon and SVM algorithm", in *National Journal of Information Technology Management*, 2015(7), pp. 345-362.

[43] S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini, "Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining,"

[56] A. Balali, A. Rajabi, S. Ghassemi, M. Asadpour, and H. Faili, "Content diffusion prediction in social networks," in *5th Conference on Information and Knowledge Technology (IKT)*, 2013, pp. 467-471.

[57] P. Turney, "Mining the web for synonyms: PMI-IR versus LSA on TOEFL," in *12th European Conference on Machine Learning (ECML 2001), Freiburg, Germany*, 2001, pp. 491-502.

[58] K. Denecke, "Using sentiwordnet for multilingual sentiment analysis," in *Data En-gineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, 2008, pp. 507-512.

[59] C. M. Özsert and A. Özgür, "Word polarity detection using a multilingual approach," in *Computational Linguistics and Intelligent Text Processing*, ed: Springer, 2013, pp. 75-82.

[60] J. Steinberger, M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, P. Lenkova*, et al.*, "Creating sentiment dictionaries via triangulation," *Decision Support Systems,* vol. 53, pp. 689-694, 2012.

[61] F. L. Cruz, J. A. Troyano, B. Pontes, and F. J. Ortega, "Building layered, multilingual sentiment lexicons at synset and lemma levels," *Expert Systems with Applications,* vol. 41, pp. 5984-5994, 2014.

[62] F. H. Mahyoub, M. A. Siddiqui, and M. Y. Dahab, "Building an Arabic Sentiment Lexicon Using Semi-Supervised Learning," *Journal of King Saud University-Computer and Information Sciences,* vol. 26, pp. 417-424, 2014.

[63] Y. Chen and S. Skiena, "Building sentiment lexicons for all major languages," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ShortPapers)* , 2014, pp. 383-389.

[64] M. Shamsfard, "Challenges and open problems in Persian text processing,"*Proceedings of LTC,* vol. 11, 2011.

[65] W. Feely, M. Manshadi, R. Frederking, and L. Levin, "The CMU METAL Farsi NLP Approach," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 4052-4055.

[66] R. Duwairi and M. El-Orfali, "A study of the effects of preprocessing strategies on sentiment analysis for Arabic text," *Journal of Information Science,* vol. 40, pp. 501-513, 2014.

[67] W. Chamlertwat, P. Bhattarakosol, T. Rungkasiri, and C. Haruechaiyasak, "Discover-ing *Language resources and linguistic theory: Typology, second language acquisition, English linguistics,* pp. 200-210, 2007.

[44] A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López, "Ranked wordnet graph for sentiment polarity classification in twitter," *Computer Speech & Language,* vol. 28, pp. 93-107, 2014.

[45] M. Montazery and H. Faili, "Automatic Persian wordnet construction," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 846-850.

[46] K. N. Lam, F. A. Tarouti, and J. Kalita, "Automatically constructing Wordnet synsets," in *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, USA*, 2014.

[47] M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi*, et al.*, "Semi automatic development of farsnet; the persian wordnet," in *Proceedings of 5th Global WordNet Conference, Mumbai, India*, 2010.

[48] P. Vossen, "A multilingual database with lexical semantic networks," *Computational Linguistics* vol. 25, pp. 628-630, 1998.

[49] F. Keyvan, H. Borjian, M. Kasheff, and C. Fellbaum, "Developing persianet: The persian wordnet," in *3rd Global wordnet conference*, 2007, pp. 315-318.

[50] A. Famian and D. Aghajaney, "Towards Building a WordNet for Persian Adjectives," *International Journal of lexicography,* pp. 307-308, 2006.

[51] M. Fadaee, H. Ghader, H. Faili, and A. Shakery, "Automatic WordNet Construction Using Markov Chain Monte Carlo," *Polibits,* pp. 13-22, 2013.

[52] N. Taghizadeh and H. Faili, "Automatic Wordnet Development for Low-resource Languages using Cross-lingual WSD," *Journal of Artificial Intelligence Research,* vol. 56, pp. 61-87, 2016.

[53] F. Mahdisoltani, J. Biega, and F. Suchanek, "YAGO3: A knowledge base from multilingual Wikipedias," in *7th Biennial Conference on Innovative Data Systems Research*, 2014.

[54] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A standard Persian text collection," *Knowledge-Based Systems,* vol. 22, pp. 382-387, 2009.

[55] H. Eghbalzadeh, B. Hosseini, S. Khadivi, and A. Khodabakhsh, "Persica: A Persian corpus for multi-purpose text mining and Natural language processing," in *Telecommunications (IST), 2012 Sixth International Symposium on*, 2012, pp. 1207-1214.

Consumer Insight from Twitter via Sentiment Analysis," *J. UCS,* vol. 18, pp. 973-992, 2012.

[68] M.-T. Martín-Valdivia, E. Martínez-Cámara, J.-M. Perea-Ortega, and L. A. Ureña-López, "Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches," *Expert Systems with Applications,* vol. 40, pp. 3934-3942, 2013.

[69] K. Denecke, "Are SentiWordNet scores suited for multi-domain sentiment classification?," presented at the Fourth International Conference on Digital Information Management, (ICDIM 2009), 2009.

احسان عسکریان دانشجوی دکترای نرم‌افزار دانشگاه فردوسی است. وی مدرک کارشناسی ارشد خود را در گرایش نرم‌افزار از دانشگاه صنعتی شریف دریافت کرد. زمینه پژوهشی مورد علاقه‌ایشان نظرکاوی، پردازش متن و داده‌کاوی است. ایشان از سال ۱۳۹۰ با عضویت در آزمایشگاه فناوری وب در پروژه‌های مختلف داده‌کاوی و پردازش متن مشارکت داشته و ابزارهای پایه‌ای متنوعی در رابطه با پردازش متن ایجاد کرده است.
نشانی رایانامهٔ ایشان عبارت است از:
ehsan.asgarian@mail.um.ac.ir

محسن کاهانی استاد گروه مهندسی کامپیوتر دانشگاه فردوسی مشهد و مدیر آزمایشگاه فناوری وب است. ایشان دکترای خود را در رشتهٔ مهندسی کامپیوتر از دانشگاه ولونگونگ استرالیا در سال ۱۳۷۷ اخذ کرده است. زمینه‌های پژوهشی مورد علاقه ایشان شامل وب‌معنایی، پردازش زبان طبیعی، سامانه‌های تصمیم‌یار و مهندسی نرم‌افزار است.
نشانی رایانامهٔ ایشان عبارت است از:
kahani@um.ac.ir

شهلا شریفی مدرک کارشناسی ارشد و دکترای خود را در رشته زبان‌شناسی همگانی از دانشگاه فردوسی مشهد دریافت کرده و در حال حاضر با درجه دانشیاری عضو هیئت علمی گروه زبان‌شناسی همین دانشگاه است. ایشان تاکنون مقالات متعدد در زمینه‌های مورد علاقه‌اش همچون کاربردشناسی زبان، ساخت واژه و رده‌شناسی زبان نگاشته که در مجلات علمی پژوهشی داخلی و خارجی یا مجموعه مقالات همایش‌ها منتشر شده‌اند و رساله‌ها و پایان‌نامه‌های متعددی را در دانشگاه فردوسی مشهد و سایر دانشگاه‌های کشور هدایت و راهنمایی کرده است.
نشانی رایانامهٔ ایشان عبارت است از:
sh-sharifi@um.ac.ir