یک چارچوب نیمه‌نظارتی مبتنی بر لغت‌نامه وفقی خودساخت جهت تحلیل نظرات فارسی

محسن نجف زاده[۱]، سعید راحتی قوچانی[۲]* و رضا قائمی[۳]

[۱] گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

[۲] گروه مهندسی برق، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

[۳] گروه مهندسی کامپیوتر، واحد قوچان، دانشگاه آزاد اسلامی، قوچان، ایران

چکیده

با معرفی وب ۲,۰ و ۳,۰ تعاملات کاربران در فضای مجازی، منجر به ایجاد انبوهی از نظرات ارزشمند شده است. با توجه به‌دشواری یا عدم امکان تحلیل و بررسی دستی این نظرات، تحلیل احساس متن و یا نظرکاوی به‌عنوان یکی از زیرمجموعه‌های پردازش زبان طبیعی مطرح شد. تلاش‌های محدودی در نظرکاوی فارسی نسبت به سایر زبان‌ها صورت گرفته است. در این مقاله برای نخستین بار، یک چارچوب نیمه‌نظارتی برای نظرکاوی فارسی ارائه شده است. درضمن، ازآنجاکه یکی از آخرین پیشرفت‌های علمی در نظرکاوی زبان فارسی الگوریتمی بر اساس استخراج الگوهای حسی وفقی (حساس به مجموعه‌داده) مبتنی بر خبره انسانی است، در این پژوهش ضمن ارتقای الگوریتم یادشده، تعیین برچسب‌های حاوی احساس به کمک یک لغت‌نامهٔ خودساخت (بدون نیاز به خبره انسانی) وفقی انجام می‌گیرد؛ همچنین کاربرد دسته‌بند مدل مخفی مارکوف خودناظر بر روی خصیصه‌های یادشده در کنار قوانین مبتنی بر معیار شباهت برای فرآیند نظرکاوی بررسی شده‌است. در راستای خودآموزسازی هوشمند، روشی برای ارزیابی قابلیت اطمینان بالای خروجی، ارائه شده‌است که خودآموزی به‌شرط وجود آن انجام می‌پذیرد. روش پیشنهادی با اجرا بر روی دادگان مبنا نرخ صحت نود درصد (باوجود عدم نیاز به خبرهٔ انسانی) را که در مقایسه با روش‌های نظارتی و نیمه‌نظارتی مستقل از خبرهٔ موجود برتری قابل‌ملاحظه‌ای دارد، خروجی می‌دهد؛ همچنین این الگوریتم نیمه‌نظارتی هنگام استفاده از مجموعه آموزش کوچک با نسبت مجموعه دادگان آموزش/آزمون ده به نود نیز بررسی و با نرخ صحت ۸۰٪ قابلیت اطمینان آن به اثبات رسید.

واژگان کلیدی: نظرکاوی، یادگیری خودناظر، لغت‌نامهٔ خودساخت، مدل مخفی مارکوف، لغت‌نامهٔ وفقی

A Semi-supervised Framework Based on Self-constructed Adaptive Lexicon for Persian Sentiment Analysis

Mohsen Najafzadeh[1], Saeed Rahati Quchani[2*] & Reza Ghaemi[3]

[1]Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

[2]Department of Electronic Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

[3]Department of Computer Engineering, Quchan Branch, Islamic Azad University, Quchan, Iran

Abstract

With the appearance of Web 2.0 and 3.0, users' contribution to WWW has created a huge amount of valuable expressed opinions. Considering the difficulty or impossibility of manually analyzing such big data, sentiment analysis, as a branch of natural language processing, has been highly considered. Despite the other (popular) languages, a limited number of research studies have been conducted in Persian sentiment analysis. In this

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

study, for the first time, a semi-supervised framework is proposed for Persian sentiment analysis. Moreover, considering that one of the most recent studies in Persian, is an algorithm based on extracting adaptive (dataset-sensitive) expert-based emotional patterns. In this research, extraction of the same state-of-the-art emotional patterns is proposed to be performed automatically. Moreover, application of the HMM classifier, by utilizing the mentioned features (as its states) is analyzed; and additionally, HMM-based sentiment analysis is upgraded by being combined with a rule-based classifier for the opinion assignment process. In addition, toward intelligent self-training, a criterion for evaluating, the high reliability of output is presented by which (assuming satisfaction of the criterion) the self-training process is performed in "lexicon-extraction" and "classifier," as learning systems. The proposed method, by being applied on the basis dataset, provides 90% of accuracy (despite its expert-independent lexicon generation nature), which in comparison with the supervised and semi-supervised methods in the state-of-the-art has a considerable superiority. Moreover, this semi-supervised method is evaluated by a 10/90 ratio of train/ test and its reliability is demonstrated by providing 80% of accuracy.

Keywords: Opinion Mining, Self-training, Self-constructed Lexicon, Hidden Markov Model, Adaptive Dictionary

۱- مقدمه

در سال‌های اخیر با معرفی وب ۲٫۰ و ۳٫۰ که بر پایهٔ مشارکت‌ها و تعاملات استوار است، شاهد گسترش فناوری‌های وب، رونق رسانه‌های اجتماعی و افزایش تعاملات کاربران در اغلب وب‌سایت‌ها هستیم که حجم زیادی از نظرات غنی و همچنین ارزان را ایجاد کرده‌اند. درصد قابل‌توجهی از این داده‌ها به‌صورت متن و صورت‌های دیگر رسانه‌ای نظیر صدا و تصویر نگهداری می‌شوند؛ اما به‌دلیل نبود یک استاندارد همه‌جانبه و دقیق در تنظیم متون و ثبت آن‌ها، این داده‌ها طبیعتی غیرساخت‌یافته و یا نیمه‌ساخت‌یافته دارند [10]. به‌طورکلی اطلاعات موجود در اسناد متنی را به دو دستهٔ عینی[1] و ذهنی[2] می‌توان تقسیم‌بندی کرد. دستهٔ عینی شامل واقعیت‌ها، دستورهای واقعی و قابل‌مشاهده دربارهٔ موجودیت‌های مستقل و اتفاقاتی است که در جهان می‌افتد؛ اما دستهٔ ذهنی بازتاب عواطف انسانی و یا مشاهداتی است که مردم نسبت به دنیای خارج و اتفاقات آن دارند [12].

در این میان، تحلیل و بررسی دستی اطلاعات ذهنی بسیار موردعلاقه و توجه بنگاه‌های خرد و کلان اقتصادی، سیاست‌مداران و ... است. تحلیل نظرات مشتریان اهمیت زیادی برای صاحبان تجارت جهت گرفتن بازخورد از مشتریان و مصرف‌کنندگان کالاها یا خدمات ارائه‌شدهٔشان دارد [9, 14]. همچنین بررسی افکار عمومی در بین منابع موجود در وب (همچون انجمن‌ها، شبکه‌های اجتماعی و ...) در مورد یک موضوع خاص مانند انتخابات [20] و یا بورس سهام [7] برای سیاست‌مداران ارزشمند است. به‌عنوان مثالی دیگر، تحلیل نقدهای یک فیلم [16] برای متصدیان این حوزه بسیار حائز اهمیت است.

اما تحلیل و بررسی این نظرات یا داده‌های ذهنی، بسیار پرهزینه و در بسیاری از موارد غیرممکن است. خوشبختانه هوش مصنوعی و به‌طور خاص پردازش زبان طبیعی[3] و متن‌کاوی[4] امکان کاوش خودکار نظرات را فراهم می‌آورد. به این فرآیند، به‌طور خاص، نظرکاوی[5] یا تحلیل احساسات متن[6] اطلاق می‌شود که با نام‌های اندیشه‌کاوی یا عقیده‌کاوی [2] نیز شناخته شده‌است و تاکنون توجه بسیاری از پژوهش‌گران را به خود جلب کرده است.

اغلب پژوهش‌های انجام‌شده در این حوزه و همچنین مجموعه‌داده‌های گردآوری شده بر روی زبان‌های انگلیسی، هلندی، چینی، اسپانیایی و ترکی است [15]. در دههٔ گذشته بیش از دو هزار پژوهش در حوزهٔ نظرکاوی انجام پذیرفته است [13] و همچنین تعداد اسناد نمایه‌شده در این حوزه، توسط پایگاه دانش اسکوپوس تا پایان سال ۲۰۱۶ میلادی به حدود شش‌هزار پژوهش می‌رسد. در زبان فارسی، تلاش‌های پژوهشی به‌نسبه اندکی در زمینهٔ نظرکاوی و همچنین ایجاد و گردآوری مجموعه‌داده‌های استاندارد در این زمینه صورت گرفته است.

از نگاهی دیگر، پژوهش‌های صورت‌گرفته برای حل مسئله نظرکاوی به سه دستهٔ کلی نظارتی، نیمه‌نظارتی و بدون‌نظارت [15] تقسیم‌بندی می‌شوند. اغلب پژوهش‌های نظرکاوی به روش نظارتی از الگوریتم‌های یادگیری ماشین بردار پشتیبان [4]، بیز ساده [17] و مدل مخفی مارکوف [24] بهره جسته‌اند. روش‌های نظارتی، مناسب برای داده‌گان‌هایی هستند که در آن به میزان کافی نظرات برچسب‌گذاری‌شدهٔ احساسی از کاربران برای آموزش یک سامانهٔ یادگیری ماشین موجود باشد؛ روش‌های نیمه‌نظارتی که دقت کمتری از آن‌ها نسبت به دستهٔ قبلی انتظار می‌رود، مناسب مسائلی هستند که داده‌های برچسب‌گذاری‌شدهٔ اندکی که فقط برای راه‌اندازی

[1] Objective
[2] Subjective

[3] Natural Language Processing (NLP)
[4] Text mining
[5] Opinion mining
[6] Sentiment analysis

اولیهٔ یک سامانه به کار می‌آیند، وجود دارد و پس‌ازآن، نظرکاو باید خودش را بهبود دهد. روش‌های بدون‌نظارت که دقت کمتر و یا سرعت کندتر نسبت به دو روش گذشته از آن‌ها انتظار می‌رود، مناسب برای مسائلی هستند که تنها نظرات بدون برچسب در آن‌ها وجود داشته باشد و چنین سامانه‌هایی به‌طورطبیعی داده‌ای برای یادگیری از بیرون ندارند.

تمرکز این مقاله بر روی دستهٔ دوم یعنی روش‌های نیمه‌نظارتی است که پژوهش‌های زیادی در این حوزه انجام‌شده است؛ ازجمله یکی از جدیدترین آن‌ها پژوهش *داسیلوا و همکاران* (۲۰۱۶) [22] در زبان انگلیسی است. اگرچه، در زبان فارسی، بنابر آخرین اطلاعات ما، پژوهشی در این حوزه صورت نپذیرفته است. در این پژوهش، الگویی نیمه‌نظارتی برای تحلیل احساسات متن ارائه می‌کنیم تا علاوه‌بر نظرکاوی با دقت بالا در زبان فارسی، دادگان[1] با داده‌های برچسب‌خوردهٔ کم را نیز بتواند پشتیبانی کند.

به‌صورت خلاصه جنبه‌های نوآوری این مقاله به شرح زیر است:

- نخستین پژوهش در حوزهٔ نظرکاوی نیمه‌نظارتی در زبان فارسی (به‌طور خاص در حوزهٔ خودآموزی).
- خودکارسازی استخراج خصیصه‌های حسی به‌جای استفاده از خبرهٔ انسانی
- استفاده از چندنگاشت‌های عامیانه، به‌عنوان یک رویکرد فرااکتشافی، برای پوشش‌دادن چالش‌های بررسی‌نشدهٔ زبان فارسی
- استفاده از ویژگی‌های زوج مرتبی (برچسب‌صرفی، برچسب احساسی) به‌عنوان وضعیت‌های طبقه‌بند مدل مخفی مارکوف برای افزایش دقت نظرکاوی در زبان فارسی.

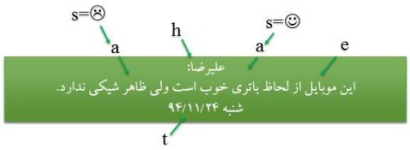
ابتدا در بخش دو مروری بر پژوهش‌های پیشین خواهیم داشت؛ بخش سوم به معرفی روش پیشنهادی اختصاص‌یافته است؛ در بخش چهارم نتایج تجربی گزارش و ارزیابی شده و درنهایت در بخش پنجم به نتیجه‌گیری و ارائه کارهای آینده خواهیم پرداخت.

۲- مروری بر پژوهش‌های پیشین

۲-۱- تعاریف پایه

نظرکاوی به استخراج نظرات کاربران و تشخیص قطبیت[2] آن‌ها درون متون ذهنی و نیز عینی می‌پردازد. اندیشه‌کاوی، تحلیل مستندات جهت استخراج نظرات، احساسات و خواسته‌های کاربران در یک حوزهٔ خاص است. در تعریفی که در [12] برای نظرکاوی ارائه می‌شود، نظرکاوی یا تحلیل احساسات، یک نظر به‌صورت زیر معرفی شده‌است:

نظر: یک نظر[3] شامل پنج‌تایی (e, a, s, h, t) است که در آن e نام موجودیت، a جنبه‌های موجودیت، s لحن نظر بر روی جنبه a از موجودیت e، h نظردهنده و t زمانی است که نظردهنده h نظر خود را در مورد موجودیت e بیان می‌کند که در شکل (۱) قابل‌مشاهده است.



(شکل-۱): نمونه‌ای از یک نظر
(Figure-1): An example from a comment

۲-۲- چالش‌ها

چالش‌های فراوانی در حوزهٔ نظرکاوی وجود دارد. ازجمله بیان غیرصریح مانند: «برای خریدش بیشتر فکر کن»، وجود طعنه، کنایه یا شوخی مانند: «گوشی ش به درد گردو شکستن می‌خوره!»، زبان گفتاری (عامیانه یا غیررسمی) مانند: «این گوشی رو یه موقع از دست ندیدش» و وجود نظرات جعلی.

در زبان فارسی نیز به سبب ویژگی‌های خاص آن و درعین‌حال نهادینه‌نشدن سبک نگارش استاندارد، با چالش‌هایی روبه‌رو هستیم. نمونه‌هایی از این موارد عبارت‌اند از: وجود یا عدم وجود تشدید «معین» و تنوین «واقعاً»، تنوع شیوه دگرنویسی (گردآوری / گرداوری) و پیوسته‌نویسی (سرهم یا با نیم‌فاصله) یا جدانویسی (کتاب‌شناسی / کتابشناسی/ کتاب‌شناسی) و ... [3].

از دیگر چالش‌های این حوزه، به هزینه‌بر بودن تهیهٔ داده‌های برچسب دار در روش‌های نظارتی می‌توان اشاره کرد. این چالش خود یکی از دلایل پدیدارشدن روش‌های نیمه‌نظارتی؛ است که هم از مزایای هوشمندی یک سامانه بدون‌نظارت و هم از مزایای وجود تعداد کمی دادهٔ برچسب‌دار برای راه‌اندازی و آموزش اولیه بهره می‌جوید.

۲-۳- رویکردهای حل مسئله

به‌طورکلی برای حل مسئله نظرکاوی سه رویکرد استفاده از لغت‌نامه، استفاده از الگوریتم‌های یادگیری ماشین و رویکرد ترکیبی ارائه شده‌است [۱۵] که به بررسی آن‌ها خواهیم پرداخت.

[1] Dataset
[2] Polarity
[3] Opinion

۱-۳-۲- رویکرد استفاده از لغت‌نامه

در رویکرد استفاده از لغت‌نامه، از منابع کمکی برای طبقه‌بندی استفاده می‌شود این منابع می‌توانند یک فرهنگ لغت، اطلاعات فراهم‌شده توسط یک موتور جستجوگر یا مجموعه‌ای از صفات و قیود باشد. یکی از مدل‌های متنی متداول در این دسته از روش‌ها، استفاده از الگوهای ثابت مستخرج از برچسب جملات مانند جدول (۱) [11] است.

(جدول-۱): الگوهای ثابت مستخرج جهت نظرکاوی [11]
(Tabel-1): Fixed patterns, extracted for Opinion Mining [11]

واژه اول	واژه دوم	واژه سوم
JJ	NN or NNS	سایر
RB,RBR, or RBS	JJ	not NN n or NNS
JJ	JJ	not NN n or NNS
NN or NNS		not NN n or NNS
RB,RBR, or RBS	VB,VBD,VBN, *or VBG*	سایر

JJ - Adjective
NN - Noun, singular or mass
NNS - Noun, plural
NNP - Proper noun, singular
VB - verb. base form
VBD - Verb, past tense
VBG - Verb, gerund or present participle
RB - Adverb
RBR - Adverb, comparative
RBS - Adverb, superlative

به نظر می‌رسد نخستین کار حوزه نظرکاوی فارسی را *شمس، شاکری و فیلی* [18] ارائه دادند. در این مقاله از لغت‌نامه‌ای شامل ۸۰۲۷ کلمه استفاده شده که در اصل به زبان انگلیسی بوده و نویسندگان از ترجمه فارسی آن بهره جسته‌اند. در آن پژوهش دقت حدود هشتاد درصد برای طبقه‌بندی نظرات گزارش شده‌است. *بصیری، نقشنیلچی و آقایی* [6] نیز چارچوبی برای نظرکاوی در زبان فارسی با استفاده از یادگیری غیرنظارتی ارائه کرده و با استفاده از آن ده درصد بهبودِ دقت را در مقایسه با روش مطرح‌شده در [۱۷] (که در بخش بعد به آن اشاره خواهیم نمود) بر مجموعه‌داده‌ای با نام BG Data گزارش کردند.

۲-۳-۲- رویکرد اســتفاده از الگوریتم‌های یادگیری ماشین

دستهٔ دوم روش‌های حل مسئله نظرکاوی از الگوریتم‌های یادگیری ماشین و دسته‌بندهای مختلف مبتنی بر یادگیری ماشین استفاده می‌کنند. فرآیند یادگیری در طبقه‌بندهای استفاده‌شده در این دسته، با استفاده از ویژگی‌های متون و نظرات (مانند TF-IDF[1]) و درصورتی‌که یادگیری تحت نظارت[2] باشد با استفاده از برچسب، نظر مربوط به هر متن انجام می‌شود.

سارایی و باقری [17] با تمرکز بر انتخاب ویژگی در نظرکاوی فارسی، روش نوینی برای نظرکاوی فارسی ارائه دادند که در آن (بعد از ریشه‌یابی کلمات) از ویژگی‌هایی مثل تکرار کلمات، واریانس فرکانس کلمه[3] و ویژگی ارائه‌شده توسط اطلاعات متقابل MI[4]) استفاده شده و بعد از پالایش ویژگی‌ها به‌وسیله الگوریتم یادگیری ماشین بیز ساده(NB)[5] مرحله آموزش ویژگی ایجاد مدل آغاز می‌شود؛ نویسندگان با پیاده‌سازی این روش بر روی داده‌گانی که با ۸۲۹ نظر فارسی در خصوص محصولات مختلف تلفن همراه گردآوری کردند، دقت ۸۵ درصد را ارائه دادند.

روش نیمه‌نظارتی در حوزهٔ نظرکاوی نخستین بار در سال ۲۰۰۹ توسط *رو و راویچاندران* [26] معرفی شد و بهترین نتایج یافت‌شده توسط *بیکر، ارهارت و همکارانشان* [27] در سال ۲۰۱۳ با الگوریتم یادگیری Self-training و معیار دقت ۰/۶۴۱ حاصل شده‌است. تعدادی از پژوهش‌های نیمه‌نظارتی انجام‌شده در زبان انگلیسی در جدول (۲) آورده شده است.

۳-۳-۲- رویکرد ترکیبی

دستهٔ سوم روش‌های حل مسئله نظرکاوی، روش‌های هر دو دستهٔ یادشدهٔ قبلی را هم‌زمان و به‌صورت ترکیبی استفاده می‌کند. ایده اصلی روش‌های ترکیبی به‌کارگیری هم‌زمان مزایای رویکرد استفاده از لغت‌نامه مانند سرعت و قدرت پردازش و نیز مزایای رویکرد استفاده از یادگیری ماشین ازجمله عدم وابستگی به ساختار دستور زبانی جمله است.

علیمردانی و آقایی [4] روشی ترکیبی برای نظرکاوی فارسی ارائه کرده‌اند. در این روش ابتدا معادل انگلیسی کلمات به‌وسیله لغت‌نامه SentiWordNet جستجو شده و سپس بر اساس یافتن معادل واژه اصلی، مثبت، منفی و یا خنثی‌بودن قطبیت آن گزارش می‌شود و سپس این لغات به‌عنوان ویژگی توسط الگوریتم ماشین بردار پشتیبان[6] برای طبقه‌بندی نظرات مورداستفاده قرار می‌گیرد؛ بهترین نتیجهٔ گزارش‌شده برای طبقه‌بندی نظراتی که توسط ایشان از داده‌های مربوط به یک تارنمای هتل گردآوردی شده بود، ۸۳/۵٪ بوده است.

یکی از جدیدترین پژوهش‌های انجام‌شده در این حوزه،

[1] Term Frequency–Inverse Document Frequency (TF-IDF)
[2] Supervised Learning
[3] Term Frequency Variance (TFV)
[4] Mutual information (MI)
[5] Naive Bayes (NB)
[6] Support vector machine (SVM)

روشی است *که آقایان اصغری، کاهانی و عسگریان* [1] ارائه کرده‌اند. روش ایشان با استفاده از نقش‌های صرفی کلمات و نیز تعریف برچسب‌های خاص احساسی و ترکیب آن‌ها با یکدیگر، الگوهای کامل‌تری را از بیان حس در جملات نظری معرفی می‌کند؛ اما ازآنجاکه این الگوها با توجه به نظرات کاربران در بخش آموزش و به‌صورت خبره‌آورد (مبتنی بر خبره) استخراج می‌شود، این روش تنها بعد از ایجاد لغت‌نامه احساسی خبره‌آوردِ بومی برای یک مجموعه‌داده، تنها بر روی همان مجموعه‌داده قابلیت استفاده خواهد داشت. در عوض، روش ارائه‌شده با آزمایش بر روی مجموعه‌دادهٔ آزمون ارائه‌شده توسط آزمایشگاه فناوری وب[1]، نرخ بالای صحت ۹۴٪ را در پی داشت.

(جدول-۲): نگاهی بر چند الگوریتم متداول در روش‌های نیمه‌نظارتی نظرکاوی توییت‌های انگلیسی [25]

(Tabel-2): Overview of literature on SSL for tweet sentiment analysis [25]

نام روش نیمه‌نظارتی	مجموعه‌داده	*F-score*
Self-training [26]	SemEval 2013	0.641
Self-training [27]	SemEval 2013	0.637
Co-training [28]	TREC 2011	در دسترس نیست
Co-training [29]	Taco Bell	نمونه‌های مختلف برآورد شده است

۳- روش پیشنهادی

روش ارائه‌شده در این مقاله در سطح جمله و با رویکرد ترکیبی است که از روش‌های زیر بهره می‌برد:

(جدول-۳): خلاصه گام‌های روش پیشنهادی

(Table-3): Steps the proposed method

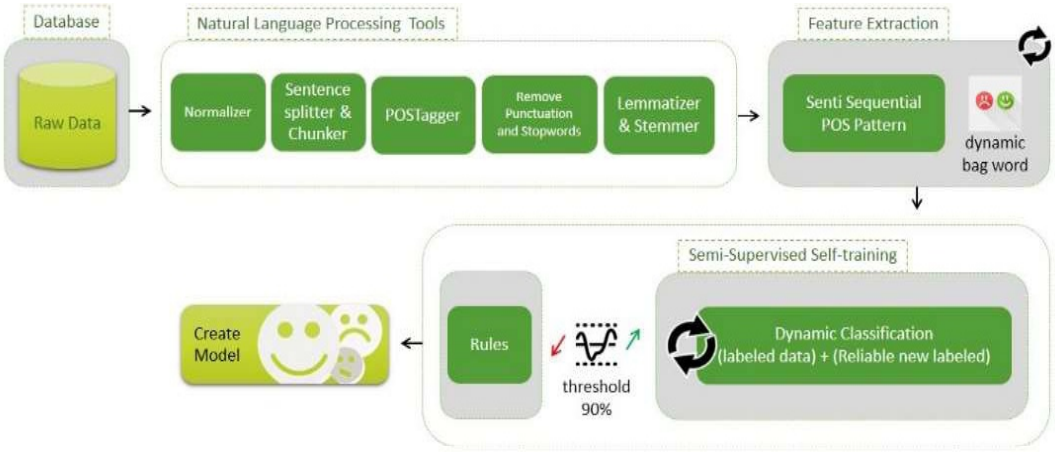
گام	هدف
آموزش	در این مرحله مجموعه‌داده آموزشی وارد برنامه شده و بدون پیش‌پردازش چندنگاشت‌ها (یک، دو و سه)ی حسی با شرایط خاص، ساخته می‌شوند و این واژه‌ها (به‌جز ایست‌واژه‌ها) و عبارات حسی ایجادشده به لغت‌نامه وفقی پویا افزوده می‌شود؛ سپس پیش‌پردازش بر روی نظر جاری انجام می‌گیرد؛ سپس برداری شامل معادل عددی برچسب‌های تخصیص‌یافته، به‌عنوان ورودی‌های الگوریتم یادگیری ماشین تشکیل می‌شود؛ مدلی بر اساس مقادیر ورودی (بردارهای متناظر با برچسب جملات) و دسته‌های تعیین‌شده، ایجاد خواهد شد.
آزمون	در این مرحله در ابتدا بدون درنظرگرفتن پیش‌پردازش، برچسب‌گذاری اجزای کلام صورت می‌گیرد. کلمات با همان برچسب و چندنگاشت‌ها با برچسب فرضی <*Phrase*>، بر مبنای معیاری از تعداد تکرار و دارابودن شرایط خاص و تشخیص مثبت و یا منفی‌بودن، برچسب آن‌ها با پیشوند *P_* و یا *N_* ویرایش می‌شود.
	ممکن است، واژگانی در مرحله قبل برچسب‌گذاری نشده باشند؛ لذا کل جمله پیش‌پردازش شده و لم‌یابی و ریشه‌یابی نیز صورت می‌گیرد و سپس برچسب‌گذاری اجزای کلام انجام می‌شود
	در مرحله بعد معادل عددی هر برچسب به‌عنوان برداری عددی ایجاد می‌شود و به‌کمک الگوریتم خودآموز HMM تشخیص دسته‌بندی انجام می‌گیرد.
	چنانچه نرخ شباهت به‌دست‌آمده از مقدار آستانه ما بیشتر بود، نتیجه حاصل از الگوریتم یادگیری‌مان مورد تأیید قرار می‌گیرد؛و در غیر این صورت دسته بر اساس قوانین سه‌گانه که در نظر گرفته‌شده است، تشخیص داده می‌شود.
پایانی	در این گام با بهره‌گیری از روش نیمه‌نظارتی خودیادگیر (*Self-training*) که در پایان هر مرحله آزمون انجام می‌گیرد، درصورت یکسان‌بودن نتایج حاصل از الگوریتم و قوانین، چندنگاشت‌های جمله جاری به لغت‌نامه وفقی پویا افزوده و بردار عددی حاصل نیز به داده‌های الگوریتم یادگیری ماشین افزوده می‌شود؛ و بازآموزی انجام و مدل جدید بازسازی خواهد شد.

به‌طور دقیق‌تر باید گفت در این مقاله امکان گسترش مجموعه آموزش با استفاده از نمونه‌های طبقه‌بندی‌شده با اطمینان بالا (خویش‌آموزی) وجود دارد و نیز کلیهٔ الگوهای حسی به‌صورت خودکار و هوشمند استخراج‌شده و نیز به‌جای تک‌نگاشت‌های استفاده‌شده در [1]، برای افزایش دقت، از چندنگاشت‌های حسی استفاده شده‌است و نیز به‌صورت پیش‌فرض از الگوریتم خودآموز HMM استفاده می‌کنیم (که از پتانسیل این داده‌های برچسب‌دار برای آموزش اولیه بهره می‌جوید) و تنها درصورتی‌که خروجی HMM ضریب اطمینان کافی را نداشته باشد از الگوریتم مبتنی بر قانون که ظرفیت کمتری برای دقت بالا دارد، استفاده می‌شود.

۱-۳- مراحل روش پیشنهادی

سامانهٔ ارائه‌شده، نظرکاوی را مطابق شکل (۲)، طی سه مرحله انجام می‌دهد که شرح آن‌ها در ادامه ذکر می‌شود.

[1] http://wtlab.um.ac.ir



(شکل-۲): مدل مفهومی روش پیشنهادی
(Figure-2): Conceptual model of the proposed method

۱-۱-۳- پیش‌پردازش

این مرحله شامل چندین گام مقدماتی است که قبل از ایجاد بردار (مرحله بعد) باید پیموده شوند. در این مرحله، جداسازی جملات، شناسایی کلمات، تبدیل جملات محاوره به رسمی (با استفاده از ابزار ارائه‌شدهٔ آزمایشگاه فناوری وب)، نرمال‌سازی، یکسان‌سازی نویسه‌ها و اصلاح نشانه‌گذاری‌ها انجام می‌شود.

۲-۱-۳- برچسب‌زنی احساسی اجزای کلام

در این مرحله (پس از حذف جداکننده‌ها، حروف ربط و ایست‌واژه‌ها) کلمات حاصل لم‌یابی و ریشه‌یابی می‌شوند و سپس به‌کمک برچسب‌زن اجزای کلام[1] در کنار لغتنامه وفقی[2] واژگان احساسی دارای برچسب قطبیت (که چگونگی ساخت آن‌ها در بخش ۲-۳- آمده است)، برچسب‌های حاوی حس را مشخص و به‌روزرسانی می‌کنیم. خروجی‌های حاصل از این مرحله، ورودی‌های دسته‌بند نیمه‌نظارتی‌مان را تشکیل خواهند داد.

۳-۱-۳- دسته‌بندی نیمه‌نظارتی و قوانین

در این مرحله به‌کمک یک دسته‌بند نیمه‌نظارتی که توسط داده‌های اولیه آموزش دیده شده‌است به پیش‌بینی نوع نظر، خواهیم پرداخت. یک دسته از روش‌های نیمه‌نظارتی[3] روش‌های خودیادگیر[4] است که به‌عنوان یکی از ایده‌دهندگان اولیه در [21] ارائه شد. ایدهٔ اصلی در این روش این است که فرآیند یادگیری می‌تواند از نتایج پیش‌بینی خود هم به‌عنوان یک کمک در یادگیری استفاده کند. به همین دلیل این روش به نام خودآموزی[5] و یا خوداندازه[6] نیز شناخته شده‌است.

```
Tain f from (X_l, Y_l)
    Predict on x ∈ X_u
    Add (x,f(x)) to labeled data
Repeat
```

(شبه‌کد-۱): روش نیمه‌نظارتی خودیادگیر
(Pseudocode-1): The Semi-supervised Self-training method

بعد از آموزش و راه‌اندازی دسته‌بند خودآموز، در مرحلهٔ آزمون، چنانچه معیار بیشترین شباهت[7] - بر اساس محتمل‌ترین توالی بررسی‌شده در این مدل- به‌دست‌آمده از دسته‌بند، از یک مقدار آستانه بیشتر باشد (به‌نحوی‌که دسته‌بند طبقه‌بندی را با اطمینان بالایی انجام داده باشد) نظر موردبررسی و برچسب پیش‌بینی‌شده‌اش را می‌پذیریم. مقدار آستانه ۹۰٪ را بر اساس بررسی نرخ صحت‌های به‌دست‌آمده از اجرای روش پیشنهادی با آستانه‌های مختلف حاصل شد که در شکل (۳) قابل‌مشاهده است. با توجه به خودآموزبودن دسته‌بندمان، خروجی منتخب را به‌شرط یکسان‌بودن با نتیجه خروجی حاصل از قوانین به ورودی‌های دسته‌بند می‌افزاییم و دوباره دسته‌بند را آموزش خواهیم داد. همچنین برچسب‌های احساسی نظر جاری را مطابق آنچه در بخش ۳-۲- آمده است، به لغت‌نامه قطبیت‌دار اضافه می‌کنیم؛ اما چنانچه خروجی

[1] Part-Of-Speech Tagging
[2] Adaptive Lexicon
[3] Semi-Supervised
[4] Self-training
[5] Self-teaching
[6] Bootstrapping
[7] Maximum Likelihood

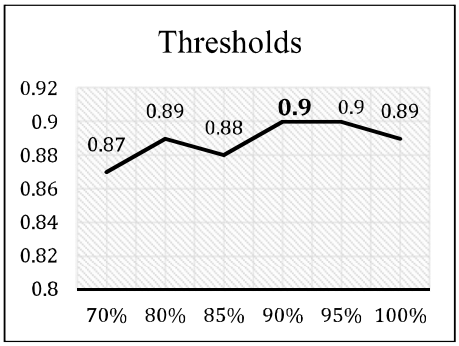
دسته‌بند از مقدار آستانه کمتر باشد (دسته‌بند طبقه‌بندی را با اطمینان کافی انجام نداده باشد)، آنگاه نتیجه حاصل از قوانین و الگوهای تعریف‌شده در جدول (۴) برای خروجی پذیرفته خواهد شد.

(جدول-۴): قوانین و الگوهای حسی (برگرفته از [1] با اعمال تغییراتی در شرط سوم)

(Table-4): Rules and emotional patterns (implied from [1] with modifications in the third rule)

شرط	عمل
چنانچه جمله حاوی برچسب حسی نباشد:	به همان نتایج دسته‌بند اعتماد می‌شود.
اگر نظر حاوی یک برچسب حسی و یا چند برچسب حسی موافق باشد:	مثبت یا منفی‌بودن آن (ها) گزارش داده می‌شود.
اگر جمله حاوی چندین برچسب ناموافق باشد:	در صورت مشاهده نفی‌کننده (غیر فعل) نخستین برچسب حسی بعد آن معکوس خواهد شد؛ در صورت مشاهده فعل منفی قطبیت کلی معکوس خواهیم کرد؛ در سایر حالات بسته به نوع واژه و یا عبارت، قطبیت منفی و یا مثبت در نظر گرفته می‌شود؛ درنهایت مقدار کلی را گزارش خواهیم کرد.

در ادامهٔ این بخش، مدل زبانی پیشنهادی بر اساس لغت‌نامه وفقی و همچنین روش‌های دسته‌بند استفاده‌شده به‌طور مبسوط شرح داده خواهد شد.



(شکل-۳): انتخاب بهترین آستانه بر اساس نرخ صحت به‌دست‌آمده از آستانه‌های متفاوت

(Figure-3): Choose the best threshold based on the accuracy obtained from different thresholds

۲-۳- لغت‌نامه وفقی و مدل زبانی پیشنهادی

قبل از معرفی مدل زبانی پیشنهادی، لازم است روش ایجاد لغت‌نامه وفقی استفاده‌شده را شرح دهیم: در مرحله آموزش به‌کمک مجموعه‌داده برچسب‌خورده مثبت و منفی، سه ویژگی تک‌واژه (تک‌نگاشت)[1]، دو واژه (دونگاشت)[2] و سه‌واژه (سه‌نگاشتی)[3] را با این شرط که به کلمات بازدارنده یا ایست‌واژه‌ها[4] ختم نشود، استخراج می‌کنیم.

لازم به ذکر است در مرحلهٔ آموزش برای ازدست‌ندادن کلمات یا چندنگاشت‌های محاوره‌ای افراد، از طرفی و از سوی دیگر، نظر به تنوع بسیار بالای کلمات محاوره‌ای در زبان فارسی، تبدیل محاوره به رسمی انجام نمی‌پذیرد؛ در مرحلهٔ آزمون نیز، در ابتدا عبارت ورودی، بدون انجام محاوره به رسمی، (کلمه به کلمه) نخست به‌صورت یک سه‌نگاشت در میان سه‌نگاشت‌های بررسی‌شده در مرحلهٔ آزمون بررسی می‌شود؛ در صورت عدم وجود به‌صورت دونگاشت بررسی‌شده و درصورت عدم وجود به‌صورت تک‌نگاشت بررسی می‌شود. اگر در هیچ‌کدام از این حالات چندنگاشت موردنظر در میان چندنگاشت‌های موجود در مرحلهٔ آموزش یافت نشد، آنگاه تبدیل محاوره به رسمی انجام می‌پذیرد تا چنانچه شکل رسمی این چندنگاشت در فاز آزمون دیده شد، از قلم نیفتد.

ازآنجاکه ذاتاً بسیاری از چندنگاشت‌ها بدون ایست‌واژه ماهیت خود را از دست می‌دهند، در قسمت بررسی سه‌نگاشت و دونگاشت‌ها (که قبل از بررسی تک‌نگاشت‌ها انجام می‌پذیرد) ایست‌واژه‌ها هنوز در متن وجود دارند؛ اما پس از رسیدن به بررسی تک‌نگاشت‌ها ایست‌واژه‌ها (که خودشان نیز تک‌نگاشت هستند) نادیده گرفته می‌شوند.

(جدول-۵): نمونه‌هایی از چندنگاشت‌های تولیدشده در لغت‌نامه وفقی

(Table-5): Examples of the generated n-grams in the adaptive dictionary

تک‌نگاشت	هوشمندانه، عالی، بی نظیریه، فوق العادس، مزخرف
دونگاشت	خوب هست، فاصله مناسب، ساده بود، کاهش کیفیت، راضی کننده، محشرهست، کم ندارد، رزولوشن پایین، *Ram* پایین، خیلی شیک
سه‌نگاشت	ارزشش را دارد، اصلا مشکلی ندارد، ولی رنگ صفحه، منحصر به فرد، سخت افزار قوی

همان‌طور که در جدول (۵) تعدادی از چندنگاشت‌های تولیدشده را مشاهده می‌کنید، مواردی منحصربه‌فرد و به‌نوعی

[1] Unigram
[2] Bigram
[3] Trigram
[4] Stop words

مبتنی بر نوع اطلاعات مجموعه‌دادهٔ مورداستفاده ایجادشده است که قطعاً بسیاری از این چندنگاشت‌ها و حتی کلمات در لغت‌نامه‌های معتبر وجود ندارند.

لذا مدل زبانیِ موردنیاز برای حل مسئله به این صورت است که در ابتدا وضعیت‌ها را به‌صورت یک دوتایی (برچسب‌صرفی، احساس) در نظر می‌گیریم که برچسب‌صرفی یکی از حالات محدود فعل، اسم، صفت، قید ... بوده و احساس نیز یکی از سه حالت مثبت، منفی و خنثی است. تمامی دونگاشت و سه‌نگاشت‌ها برای پیش‌گیری از زیادشدن تعداد وضعیت‌ها از برچسب صرفی‌ای به نام «چندنگاشت» بهره می‌جوییم. درواقع، تحلیل هر چندنگاشت به‌طور مجزا (در مرحلهٔ آموزش و آزمون) تنها به‌دلیل شناسایی و حصول اطمینان نسبت به برچسب احساس آن است و چندنگاشت‌ها وضعیت‌های جداگانه‌ای ایجاد نمی‌کنند و سپس با بررسی چندنگاشتی‌ها و کلمات در لغت‌نامه وفقی مشابه شبه‌کد ۲، برچسب را بر اساس نسبت جمع برداری منفی/مثبت‌ها بر جمع اسکالر آن‌ها، با علامتی مثبت، منفی و یا خنثی علامت‌گذاری می‌کنیم. در همین بخش منفی‌کننده‌ها[1] بر اساس کلمه ماقبل موردبررسی قرار می‌گیرند و در صورت وجودشان برچسب موجود در دادگان معکوس می‌شود. (افعال منفی که با حرف «ن» شروع می‌شوند، ولی ریشه آن با غیر «ن» است و واژه‌های «بی»، «بدون» و «نا»، منفی‌کننده امتیاز خواهند بود).

```
Input = myNGram, myNGramTag, previousWord
        pos,neg=
        findPolarityFromDics(myNGram);
        score = (pos - neg) / (pos + neg);
        if (previousWord in Negation)
            score *=-1;
        if(score>0.2 and score<0.6)
            myNGramTag = "Pos"+myNGram;
        else if(score<-0.2 and score>-0.6)
            myNGramTag = "Neg"+myNGram;
        else
            myNGramTag = "Neu"+myNGram;
Output = myNGramTag
```

(شبه کد-۲): الگوریتم تعیین قطبیت چندنگاشت بر اساس تعداد تکرار در لغت‌نامه وفقی

(Pseudocode-2): Algorithm of determining n-grams polarity, based on observations frequency in the Adaptive Dictionary

بر اساس تحلیلی که روی چندنگاشت‌ها صورت دادیم، به‌طورمعمول با دو دستهٔ کلی از چندنگاشت‌های خنثی مواجه می‌شویم.

دستهٔ نخست چندنگاشت‌هایی هستند که بسیار تکرار شده‌اند اما تعداد استفاده آن‌ها در جملات مثبت و منفی سربه‌سر بوده است. در این نوع چندنگاشت‌ها رابطه $\frac{pos-neg}{pos+neg}$ عددی کوچک منفی یا مثبت، خروجی می‌دهد. چند مثال از این دسته عبارت‌اند از فعل «است»، راستی»، «شاید» و ...

دسته دوم چندنگاشت‌هایی هستند که خیلی کم (به‌عنوان مثال یک‌بار) در کل دادگان تکرار شده‌اند و به‌طور تصادفی تکرارشان در یک یا چند جمله مثبت یا منفی بوده است. در این حالات رابطه $\frac{pos-neg}{pos+neg}$ عدد بزرگ (یک به ۱ ی(۱-) خروجی می‌دهد که به نظر خیلی مثبت یا خیلی منفی می‌آید؛ اما این مثبت یا منفی‌بودنِ شدید، معنادار نیست چراکه به‌خاطر کمبود داده این اتفاق رخ داده‌است. مثال‌های این دسته عبارت‌اند از اسم شهر «اردکان»، کلمه «زادروز»، نام «داریوش» و ... .

بر اساس منطق بالا بازهٔ ۱- تا ۱ به پنج بخش تقسیم شد و بخش وسط (به‌خاطر شامل‌بودن دستهٔ نخست از خنثی‌ها) و دو بخش کناری (به‌خاطر شامل‌بودن بخش دوم از خنثی‌ها) به‌عنوان خنثی در نظر گرفته شد و دو بخش دیگر برحسب منفی یا مثبت‌بودن برچسب منفی یا مثبت دریافت کردند؛ لذا مطابق جدول (۶) برچسب‌گذاری می‌کنیم:

(جدول-۶): بازه‌های تعیین‌شده جهت برچسب‌گذاری چندنگاشت‌های موجود در لغتنامه وفقی

(Table-6): The determined ranges for labeling n-grams in the Adaptive Dictionary

عنوان	بازه	برچسب
مقادیر مرزی با امتیاز خیلی مثبت	[+0.6,+1]	خنثی
مقادیر مرزی با امتیاز خیلی منفی	[-1,-0.6]	
مقادیر نزدیک به صفر	[0.2,+0.2]	
مقادیر میانی مثبت	[+0.2,+0.6]	مثبت
مقادیر میانی منفی	[-0.6,-0.2]	منفی

بعد از ساخت لغت‌نامه وفقی با استفاده از داده‌های آموزش، خواهیم توانست بردار ویژگی داده‌های آزمون را بر اساس توالی برچسب‌های به‌وجودآمده تهیه کنیم.

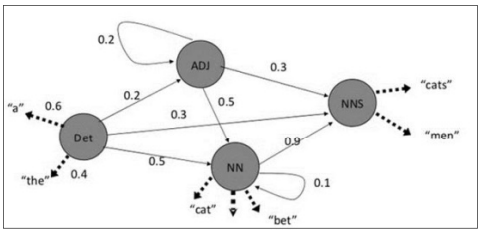
۳-۳- روش‌های دسته‌بندی استفاده‌شده

مدل مخفی مارکوف[2] در اواخر دهه ۱۹۶۰ میلادی معرفی شد. این مدل، درعمل گزینهٔ مناسبی برای حل مسائل ترتیبی بوده است؛ ازهمین‌رو در مسائل پردازش زبان طبیعی نیز مورداستفاده قرار می‌گیرد. به‌عنوان نمونه در [19] یک توسعه

[1] Negations

[2] Hidden Markov Models (HMM)

بر مدل مخفی مارکوف با استفاده از تخمین مرتبه دوم برای برچسب‌گذاری جملات مورداستفاده قرار گرفته‌است. یکی از مهم‌ترین الگوریتم‌های یادگیری بر اساس رخداد مشاهدات در مدل مخفی مارکوف، الگوریتم بام- ولش[1] است [23]؛ این روش را به‌سادگی و با محاسبه احتمال رخداد پارامترها می‌توان تعریف کرد و یکی از ویژگی‌های مخصوص این الگوریتم این است که هم‌گرایی در آن تضمین شده‌است. این دسته‌بند در جهت حل مسائل پردازش زبان طبیعی که ارتباط مستقیم با توالی کلمات دارد ازجمله برچسب‌زن اجزای کلام مورداستفاده قرارگرفته است [5].



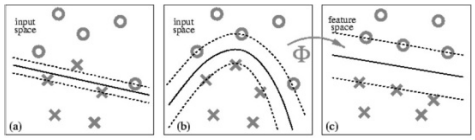
(شکل-۴): نمونه ساده از کاربرد HMM در حل مسئله POS-Tagger

(Figure-4): A simple example for HMM application in solving POS-Tagger problem

ماشین بردار پشتیبان در سال ۱۹۹۲ میلادی معرفی شد. این دسته‌بند نوع خاصی از مدل‌های خطی را می‌یابد که موجب بیشینه‌شدن تفکیک بین طبقات می‌شود. بردارهای پشتیبان درواقع نزدیک‌ترین نقاط به حاشیه ابرصفحه هستند. در مسائلی که داده‌ها به‌صورت خطی جداپذیر نباشند، یک Kernel (از نوع خطی Linear، چندجمله‌ای Polynomial، گوسی Gaussian و یا ...) برای آن در نظر گرفته می‌شود که درعمل نگاشتی را بین خط با آن ابرصفحهٔ جداکننده برقرار کند؛ یا به‌بیان‌دیگر، داده‌ها به فضای با ابعاد بیشتر نگاشت پیدا می‌کنند تا بتوان آن‌ها را در این فضای جدید به‌صورت خطی جدا کرد؛ و با این شرایط SVM را، از این منظر، غیرخطی نیز می‌توان در نظر گرفت. به‌عنوان نمونه در تصویر زیر با پارامتر φ فضای اولیه داده‌ها به فضای ویژگی‌هایی نگاشت می‌شود که با همان SVM خطی می‌توان دسته‌بندی کرد.

در این پژوهش از دسته‌بند مدل مخفی مارکوف (به‌دلیل مناسب‌بودن آن برای مشاهدات ترتیبی و مدل زبانی پیشنهادی این پژوهش) به‌عنوان دسته‌بند اصلی و از دسته‌بند ماشین بردار پشتیبان (به‌دلیل بهترین عملکرد در بین سایر دسته‌بندها مورداستفاده در سایر پژوهش‌های نظرکاوی در زبان فارسی) بهره گرفته شده‌است.

[1] Baum-Welch



(شکل-۵): شمایی از دسته‌بندی SVM مبتنی بر Kernel

(Figure-5): A schematic view of classification in Kernel-based SVM

۴-۳- معیارهای ارزیابی

جهت ارزیابی روش‌های پیشنهادی در این پژوهش، از چهار معیار دقت[2]، فراخوانی[3]، معیار F[4] و نرخ صحت[5] بر مبنای ماتریس درهم‌ریختگی[6] مطابق جدول (۷) استفاده شده‌است.

(جدول-۷): حالات ممکن صحت عمل یک دسته‌بند برای هر طبقه (TN درست منفی، TP درست مثبت، FN اشتباه منفی، FP اشتباه مثبت)

(Table-7): Possible cases for output validity of a classifier in each class (TP: true positive, TN: true negative, FP: false positive, FN: false negative)

		دسته پیش‌بینی‌شده	
		مثبت	منفی
دسته واقعی	مثبت	*True positive (TP)*	*False negative (FN)*
	منفی	*False positive (FP)*	*True negative (TN)*

معیار دقت: تعداد مواردی مثبتی که به‌درستی پیش‌بینی شده‌اند، نسبت به تعداد کل موارد درست پیش‌بینی‌شده.

$$Precision = \frac{TP}{TP + FP} \quad (۱)$$

معیار فراخوانی: تعداد مواردی مثبتی که به‌درستی پیش‌بینی شده‌اند از تعداد کل موارد مثبت.

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

معیار F: این معیار در ارتباط با معیارهای دقت و فراخوانی به‌صورت میانگین وزن‌دار، از هر دو معیار، به‌کار می‌رود. (Recall=R, Precision=P)

$$FScore = \frac{2P.R}{P + R} \quad (۳)$$

نرخ صحت: میزان مقادیر درست پیش‌بینی‌شده نسبت به تعداد کل مجموعه‌داده، بیان‌گر نرخ صحت است؛ با توجه به اینکه این معیار تنها معیاری است (از بین موارد اشاره‌شده) که

[2] Precision
[3] Recall
[4] F- Measure (FScore)
[5] Accuracy
[6] Confusion Matrix

در آن علاوه‌بر مثبت‌های درست تشخیص داده‌شده، منفی‌های درست تشخیص داده‌شده نیز ملاک ارزیابی قرار می‌گیرند، این ملاک را ملاک اصلی مقایسه قرار داده‌ایم.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (۴)$$

سه ملاک بازیابی اطلاعاتی Precision، Recall و FScore در نظرکاوی جنبه‌گرا و غیرجنبه‌گرا دو مفهوم متفاوت بااهمیت‌های متفاوت را نشان می‌دهند؛ در نظرکاوی جنبه‌گرا این مفاهیم نشان‌گر این هستند که چه میزان جنبه‌هایی که نویسنده در مورد آن صحبت کرده است، به‌درستی بازیابی شده‌اند که همواره اطلاعات ارزشمندی است و حکایت از صحت تشخیص جنبه در الگوریتم نظرکاوی دارد؛ اما در نظرکاوی غیرجنبه‌گرا که این پژوهش نیز در این حوزه تعریف‌شده، این سه معیار میزان شناسایی درست متون با برچسب نظر مثبت است. ازآنجاکه در دادگان مورد ارزیابی (دیجی‌کالا) شناسایی درست نظرات مثبت مزیتی نسبت به شناسایی درست نظرات منفی نداشتند، این معیار فقط در نمودار تشریح پایستاری درج شده‌است تا پایستاربودن تمامی ملاک‌ها (شامل ملاک‌های یادشده) را نشان دهد.

ثبات و پایداری[1]: جهت نمایش قدرت روش پیشنهادی، پایداری آن را با درنظرگرفتن اندازهٔ کوچک‌تری از مجموعه آموزش و به‌تناسب، اندازه بزرگ‌تر مجموعه آزمون موردبررسی قرار می‌دهیم.

۴- نتایج تجربی

۴-۱- معرفی مجموعه‌داده

جهت ارزیابی روش ارائه‌شده، ازآنجاکه این مقاله نسخهٔ خودکارشده (بدون نیاز به دادگان احساسی وفقی خبره آورد) پژوهش [1] است، مجموعه‌داده آن مطالعه را که توسط نویسندگان مقاله ارائه شده بود، مجموعه‌دادهٔ مرجع این پژوهش قرار دادیم. دادگان مزبور بخشی از مجموعه‌داده گردآوری‌شده توسط آزمایشگاه فناوری وب دانشگاه فردوسی از پایگاه وب دیجی‌کالا[2] است که با توجه به نوع نظردهنده (نظر خبره، کاربر فعال و کاربر عادی) شامل نظرات رسمی، نیمه‌رسمی و عامیانه است. جزئیاتی از این مجموعه‌داده به شرح زیر است:

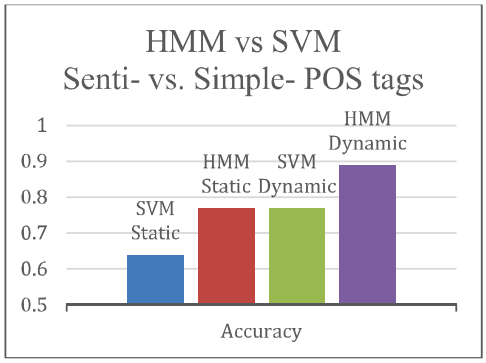
[1] Sustainability
[2] http://www.digikala.com

(جدول-۸): جزئیات مجموعه‌داده مورداستفاده
(Table-8): Details of the utilized dataset

تعداد نظرات مجموعه آموزش	*1900*
تعداد نظرات مجموعه آزمون	*229*
برچسب‌ها	مثبت و منفی
میانگین تعداد کلمات هر جمله	*8.39*
تعداد جملات مثبت	۱۶۸۴ (آموزش = ۱۵۰۷ و آزمون = ۱۷۷)
تعداد جملات منفی	۴۳۳ (آموزش = ۳۸۷ و آزمون = ۴۶)

۴-۲- نتایج

در ابتدا برای انتخاب دسته‌بند مناسب برای مدل متنی ارائه‌شده، دسته‌بند SVM را با دسته‌بند پیشنهادی (HMM) به‌کمک بردار متناظر با برچسب‌های مستخرج از دادگان احساسی ایستا واژگان فارسی دارای برچسب قطبیت تهیه‌شده در آزمایشگاه سامانه‌های هوشمند اطلاعات دانشگاه تهران توسط [8] و همچنین برچسب‌های مستخرج از دادگان احساسی پویا (وفقی) مورد ارزیابی قرار داده‌ایم؛ سپس با آزمایش بر روی دادگان نظرات مبنا، با تخصیص آموزش/آزمون به همان شیوهٔ پژوهش مورد مقایسه[3] نتایج پیش‌رو حاصل شد:



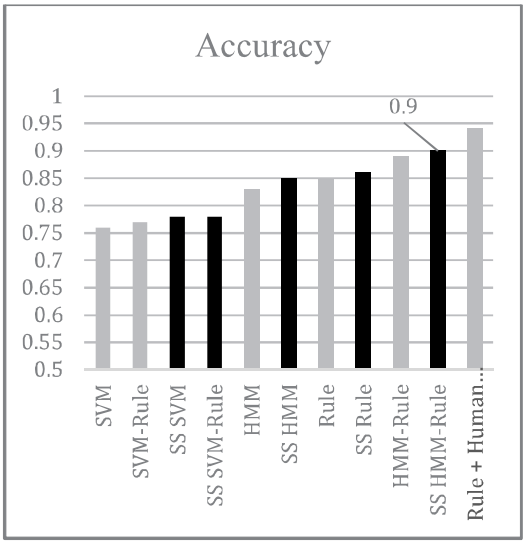
(شکل-۶): مقایسه عملکرد دسته‌بند SVM و HMM با به‌کارگیری واژه‌نامه ایستا و پویا
(Figure-6): HMM classifier vs. SVM classifier & static dictionary vs. dynamic dictionary

همان‌گونه که در شکل (۶) قابل مشاهده‌است، برچسب‌های احساسی پویا هم در مورد HMM و هم در مورد SVM نسبت به برچسب‌های احساسی ایستا نتایج را به‌طور قابل‌ملاحظه‌ای بهبود داده‌اند که نشان‌دهندهٔ کارایی برچسب‌های احساسی است. همچنین قابل‌مشاهده است که طبقه‌بند HMM در هر دو حالت، با کارایی بالاتری، نسبت به

[3] در پژوهش مورد مقایسه ۹۰ درصد داده‌ها برای آموزش و ۱۰ درصد برای آزمون به کار بسته‌شده‌اند.

طبقه‌بند SVM فرآیند نظرکاوی را انجام داده است.

در آزمایش بعد، با برگزیدن HMM به همراه برچسب‌های احساسی اثر فرآیند خودآموزی را مورد مطالعه قرار می‌دهیم. شکل (۷) کارایی طبقه‌بند HMM را یک‌بار بدون فرآیند خودآموزی و یک‌بار به همراه خودآموزی نشان داده است. مشاهده می‌شود، فرآیند خودآموزی باعث رشد تمامی مؤلفه‌های کارایی گردیده است. یکی از مزایای الگوریتم ارائه‌شده این است که در مرحلهٔ آموزش درعمل دادگان تخصصی پویا به‌صورت وفقی (سازگار با مجموعه خاصمان) ایجاد می‌شود؛ برای بررسی این مزیت، دادگان حاضر را با مجموعه دادگان حسی معتبرِ ایستا، معرفی‌شده در [8] موردبررسی و مقایسه قراردادیم. نکتهٔ قابل‌ملاحظهٔ دیگر این است که باوجود شرایط یکسان، اگرچه دقت روش پیشنهادی ما از روش مبتنی‌بر قانون و دادگان احساسی خبره آورد [1] کمتر است، اما از طرفی با این کاهش جزئی کارایی، عنصر پرهزینهٔ خبرهٔ انسانی برای استخراج دادگان احساسیِ متناسب با مجموعه‌داده، از الگوریتم حذف شده‌است و از طرف دیگر این کاهش فقط به‌اندازهٔ ۴٪ بوده که نتایج بسیار نزدیکی نسبت به حالت وجود خبره است.



(شکل-۷): مقایسه صحت نظرکاوی روش‌های مختلف اعمال‌شده بر مجموعه‌داده مبنا (روش‌های نیمه‌نظارتی پررنگ‌تر رسم شده‌اند)

(Figure-7): Comparison of the opinion mining accuracy in different methods on the benchmark dataset (semi-supervised methods are highlighted)

در انتها، برای بررسی ثبات و پایداری روش نیمه‌نظارتی ارائه‌شده را بر روی مجموعه‌داده مبنا با اندازه‌هایی متفاوتی از مجموعه آموزش-آزمون موردبررسی قراردادیم که با توجه به نتایج حاصل در شکل (۸)، ثبات و پایداری الگوریتم باوجود حجم داده آموزشی کم مشهود است.

(شکل-۸): بررسی پایداری روش نیمه‌نظارتی پیشنهادی با کاهش نسبت مجموعه آموزش

(Figure-8): Sustainability of the proposed semi-supervised method with decreasing training data ratio

۵- نتیجه‌گیری و کارهای آینده

در این مقاله، یک مدل استخراج ویژگی برای نظرکاوی مبتنی بر دادگان احساسی حساس به مجموعه‌داده (وفقی) و سپس یک طبقه‌بند مناسب برای آن (مدل مخفی مارکوف خودآموز، HMM، در کنار یک طبقه‌بند مبتنی‌بر قانون به‌عنوان دریچهٔ اطمینان HMM) ارائه شد. مقایسه‌ها نمایان‌گر برتری دادگان احساسی وفقی نسبت به دادگان احساسی ایستا و همچنین برتری طبقه‌بند HMM نسبت به طبقه‌بند ماشین بردار پشتیبانی (SVM) و همچنین بهره‌وری فرآیند خودآموزی در HMM، در فرآیند نظرکاوی است. با توجه به پیچیدگی زمانی روش HMM، نظرکاوی ارائه‌شده در این مقاله مناسب نظرکاوی برای نظرات نسبتاً کوتاه (مثل نظرات کاربران در وبگاه‌ها، یا پست‌های کاربران در میکروبلاگ‌ها) است.

نکتهٔ قابل‌توجه دیگر این پژوهش، این است که بی‌نیازشدن به عنصر خبرهٔ انسانی موردنیاز در لغتنامهٔ وفقی مورداستفاده در مرزهای دانش، دقت نظرکاوی با اختلاف جزئی ۴٪ حفظ شده‌است که این در مقایسه با روش‌های نیمه‌نظارتی و نظارتی موجود یک بهبود قابل‌ملاحظه است.

در نسخهٔ نهایی این پژوهشِ در حال انجام که به‌وسیله آن بر آن هستیم روشی حتی با کارایی بالاتر از دادگان احساسی خبره‌آورد ارائه دهیم، هم الگوریتم ایجاد دادگان احساسی و هم قانون‌های مورداستفاده در طبقه‌بندی فازی خواهند شد؛ به‌طوری‌که با پوشش عدم قطعیت موجود در سامانه بتوانیم فاصلهٔ عامل‌های کارایی روش ارائه‌شده را با روش خبره‌آورد پر کنیم و نیز جملات طولانی تقطیع جمله و به بررسی جداگانه

آن‌ها پیشنهاد می‌شود و البته در این شرایط تنها باید برچسب‌های پراهمیت‌تر در نظر گرفته شوند تا از رشد زیاد لغت‌نامهٔ پویا جلوگیری شود.

سپاس‌گزاری

در پایان از آقایان دکتر کاهانی و اصغری کمال سپاس دارم که ابزارهای پردازش زبان طبیعی آزمایشگاه فناوری وب دانشگاه فردوسی[1] را در اختیارمان قرار دادند تا به‌کمک آن در کنار کتابخانهٔ متن باز Accord.NET[2] پیاده‌سازی مقاله را انجام دهیم؛ و همچنین از *آقای دکتر اکبر زاده توتونچی* و نیز *آقای سید علی حسینی* (دانشجوی دکترای ایشان) به‌دلیل یاری و راهنمایی‌های بی‌دریغشان و همچنین از *خانم دکتر شاکری* به دلیل در اختیار دادن واژگان قطبیت‌دار فارسی سپاس‌گزاریم.

۶- References ۶- مراجع

[۱]سید محمد اصغری نکاح، محسن کاهانی و احسان. عسگریان. «نظرکاوی با استفاده از برچسب‌های صرفی و معنایی و کشف روابط حسی جملات فارسی». بیستمین کنفرانس ملی سالانه انجمن کامپیوتر ایران. دانشگاه فردوسی مشهد. ۱۳۹۴

[1] S.M. Asghari N, M. Kahani, and E. Askarian, “Opinion Mining by means of syntactic and semantic labels, and discovering emotional relations in Persian sentences”. Computer Society of Iran (20th) Computer Conference (CSICC 2015). Ferdowsi University of Mashhad. 2015

[۲]برهانی زرندی، سمیه، علی اکبر نیک نفس، و مجید محمدی. "عقیده کاوی در نقد کالا با استفاده از شبکه واژگان احساسی"، دومین کنفرانس ملی مهندسی صنایع و سیستم ها، نجف آباد، دانشگاه آزاد اسلامی واحد نجف آباد، گروه مهندسی صنایع، ۱۳۹۲

[2] S. Borhani Z., A.A. Niknafs, and M. Mohammadi, “Opinion mining in product reviews, using emotional vocabulary network” 2nd National Conference on Indurtrial Engineering & Systems (NIESC 2014). Najafabad branch, Islamic Azad University. 2014

[۳]هاجر ستوده، زهره هنرجویان. "مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تأثیرات آنها بر اثر بخشی پردازش خودکار متن و بازیابی اطلاعات". فصلنامه علمی و پژوهشی کتابداری و اطلاع رسانی - آستان قدس رضوی، ۱۳۹۱.

[3] H. Sotudeh and Z. Honarjooyan, “A review on Persian challenges in digital paradigms, and their effect on efficiency of automatic text processing and information retrieval,” Library and Information Science, 15 (4), Astan Quds Razavi. 2013

[4] S. Alimardani and A. Aghaei, "Opinion Mining in Persian Language Using Supervised Algorithms," 2015.

[5] A. Azimizadeh, M. M. Arab, and S. R. Quchani, “Persian part of speech tagger based on Hidden Markov Model,” 9th JADT, 2008.

[6] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghassem-Aghaee, “A Framework for Sentiment Analysis in Persian,” 2014.

[7] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” J. Comput. Sci., vol. 2, no. 1, pp. 1–8, 2011.

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

[9] M. Gamon, “Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis,” in Proceedings of the 20th international conference on Computational Linguistics, 2004, p. 841.

[10] V. Gupta and G. S. Lehal, “A survey of text mining techniques and applications,” J. Emerg. Technol. web Intell., vol. 1, no. 1, pp. 60–76, 2009.

[11] A. K. Jain and Y. Pandey, “Analysis and implementation of sentiment classification using lexical POS markers,” Int. J., vol. 2, no. 1, 2013.

[12] B. Liu, “Sentiment analysis and opinion mining,” Synth. Lect. Hum. Lang. Technol., vol. 5, no. 1, pp. 1–167, 2012.

[13] B. Liu, "Sentiment analysis: Mining opinions, sentiments, and emotions": Cambridge University Press, 2015.

[14] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, “Low-Quality Product Review Detection in Opinion Summarization,” in EMNLP-CoNLL, 2007, pp. 334–342.

[15] R. Kumar and R. Vadlamani, “A survey on opinion mining and sentiment analysis: tasks, approaches and applications,” Knowledge-Based Syst., vol. 89, pp. 14–46, 2015.

[16] E. Sadikov, A. Parameswaran, and P. Venetis, “Blogs as predictors of movie success,” 2009

[1] ابزارهای پردازش متون زبان فارسی، آزمایشگاه فناوری وب دانشگاه فردوسی مشهد ۱۳۹۱-۱۳۹۴ (wtlab.um.ac.ir)

[2] Accord.NET Framework, http://accord-framework.net/

[17] M. Saraee and A. Bagheri, “Feature selection methods in Persian sentiment analysis,” in Natural Language Processing and Information Systems, Springer, 2013, pp. 303–308.

[18] 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.

[19] S. M. Thede and M. P. Harper, “A second-order hidden Markov model for part-of-speech tagging,” in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 175–182

[20] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment,” ICWSM, vol. 10, pp. 178–185, 2010.

[21] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," IEEE Transactions on Information Theory, vol. 11, pp. 363-371, 1965.

[22] N. F. F. da Silva, L. F. Coletta, E. R. Hruschka, and E. R. Hruschka Jr, "Using unsupervised information to improve semi-supervised tweet sentiment classification," Information Sciences, vol. 355, pp. 348-365, 2016.

[23] L. R. Welch, "Hidden Markov models and the Baum-Welch algorithm," IEEE Information Theory Society Newsletter, vol. 53, pp. 10-13, 2003.

[24] M. Kang, J. Ahn, and K. Lee, “Opinion mining using ensemble text hidden Markov models for text classification.” 2017.

[25] N. F. F. D. Silva , L. F.Coletta, , & E. R.Hruschka, “A survey and comparative study of tweet sentiment analysis via semi-supervised learning.” ACM Computing Surveys (CSUR), 49(1), 15. 2016

[26] 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 (pp. 675-682). Association for Computational Linguistics. 2009.

[27] L. Becker, G. Erhart, D. Skiba, and V. Matula, “AVAYA: Sentiment Analysis on Twitter with Self-Training and Polarity Lexicon Expansion”. SemEval@ NAACL-HLT, pp. 333-340, 2013.

[28] S.Liu, F.Li, F.Li, X.Cheng, & H.Shen, "Adaptive co-training SVM for sentiment classification on tweets". In Proceedings of the 22nd International Conference on World Wide Web Information & Knowledge Management (pp. 2079-2088). ACM. 2013.

[29] S.Liu, W.Zhu, N.Xu, F.Li, X. Q.Cheng, Y.Liu, & Y.Wang, "Co-training and visualizing sentiment evolvement for tweet events". In Proceedings of the 22nd International Confer-ence on World Wide Web (pp. 105-106). ACM. 2013

محسن نجف زاده، کارشناسی خود را در رشته مهندسی کامپیوتر-نرم‌افزار در سال ۱۳۸۷ و مدرک کارشناسی ارشد را در دانشگاه آزاد اسلامی واحد مشهد در رشته مهندسی کامپیوتر – هوش مصنوعی در سال ۱۳۹۵ به پایان رساند. زمینه پژوهشی وی پردازش زبان طبیعی، نظرکاوی و شبکه‌های عصبی است.

نشانی رایانامهٔ ایشان عبارت است از:

mohsen.najafzadeh@mshdiau.ac.ir

سعید راحتی قوچانی، در سال ۱۳۶۸ کارشناسی خود را در رشته الکترونیک دانشگاه تهران به پایان رساند و کارشناسی ارشد و دکترای خود را در رشته مخابرات به ترتیب در واحد تهران جنوب و واحد علوم و تحقیقات دانشگاه آزاد اسلامی در سال‌های ۱۳۷۲ و ۱۳۷۷ اخذ کرد. سپس با پایان دوره دکترا، به‌عنوان استادیار و از سال ۱۳۹۰ در سمت دانشیاری دانشگاه آزاد اسلامی واحد مشهد مشغول به خدمت شد. تاکنون بیش از یک‌صد مقاله در نشریات و کنفرانس‌ها از ایشان به چاپ رسیده است. گرایش پژوهشی ایشان پردازش زبان و گفتار و آموزش شبکه‌های عصبی و کاربرد آن در مدل‌سازی سامانه‌های زیستی است.

نشانی رایانامهٔ ایشان عبارت است از:

rahati@mshdiau.ac.ir

رضا قائمی، مدرک کارشناسی خود را در رشته مهندسی کامپیوتر در سال ۱۳۷۶ و همچنین مدرک کارشناسی ارشد را در رشته مهندسی کامپیوتر-نرم‌افزار در سال ۱۳۷۹ به اتمام رساند؛ سپس مدرک دکترای خود را در سال ۱۳۹۱ در رشته هوش مصنوعی از دانشگاه یوپیام مالزی اخذ کرد. وی هم‌اکنون عضو هیئت‌علمی و استادیار دانشگاه آزاد اسلامی واحد قوچان است. زمینهٔ پژوهشی ایشان هوش محاسباتی و داده‌کاوی است.

نشانی رایانامهٔ ایشان عبارت است از:

r.ghaemi@iauq.ac.ir