یک مدل موضوعی احتمالاتی مبتنی بر روابط محلّی واژگان در پنجره‌های هم‌پوشان

مرضیه رحیمی*، مرتضی زاهدی و هدی مشایخی

دانشکدهٔ مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی شاهرود، شاهرود، ایران

چکیده

بسیاری از مدل‌های موضوعی مانند LDA که مبتنی بر هم‌رخدادی واژگان در سطح یک سند هستند قادر به بهره‌گیری از روابط محلی واژگان نیستند. برخی از مدل‌های موضوعی مانند BTM سعی کرده‌اند با ترکیب موضوعات و مدل‌های زبانی n-gram، این مشکل را حل کنند. امّا BTM مبتنی بر ترتیب دقیق واژگان است؛ بنابراین با مشکل تُنُکی روبه‌روست. در این مقاله یک مدل موضوعی احتمالاتی جدید معرفی شده که قادر به مدل‌کردن روابط محلی واژگان با استفاده از پنجره‌های هم‌پوشان است. بر اساس فرضیه هم‌رخدادی، رخداد هم‌زمان واژگان در پنجره‌های کوتاه‌تر، گواه محکم‌تری بر ارتباط معنایی آنهاست. در مدل پیشنهادی، هر سند، مجموعه‌ای از پنجره‌های هم‌پوشان فرض می‌شود، که هریک متناظر با یکی از واژگان متن است. موضوعات بر مبنای هم‌رخدادی واژگان در این پنجره‌های هم‌پوشان استخراج می‌شوند. به‌عبارت دیگر، مدل پیشنهادی، روابط محلی واژگان را بدون وابستگی به ترتیب دقیق آنها مدل می‌کند. آزمایش‌های ما نشان می‌دهد که روش پیشنهادی، موضوعات منسجم‌تری را تولید و در کاربرد خوشه‌بندی اسناد، دقیق‌تر از دو مدل LDA و BTM عمل می‌کند.

واژگان کلیدی: مدل‌های موضوعی احتمالاتی، نمونه‌برداری گیبس، هم‌رخدادی، مدل‌های گرافیکی، خوشه‌بندی متن

A Probabilistic Topic Model based on Local Word Relationships in Overlapped Windows

Marziea Rahimi, Morteza Zahedi & Hoda Mashayekhi

Department of Computer Engineering, Shahrood University of Technology,Shahrood, Iran

Abstract

A probabilistic topic model assumes that documents are generated through a process involving topics and then tries to reverse this process, given the documents and extract topics. A topic is usually assumed to be a distribution over words. LDA is one of the first and most popular topic models introduced so far. In the document generation process assumed by LDA, each document is a distribution over topics and each word in the document is sampled from a chosen topic of that distribution. It assumes that a document is a bag of words and ignores the order of the words. Probabilistic topic models such as LDA which extract the topics based on documents-level word co-occurrences are not equipped to benefit from local word relationships. This problem is addressed by combining topics and n-grams, in models like Bigram Topic Model (BTM). BTM modifies the document generation process slightly by assuming that there are several different distributions of words for each topic, each of which correspond to a vocabulary word. Each word in a document is sampled from one of the distributions of its selected topic. The distribution is determined by its previous word. So BTM relies on exact word orders to extract local word relationships and thus is challenged by sparseness. Another way to solve the problem is to break each document into smaller parts for example paragraphs and use LDA on these parts to extract more local word relationships in these small parts. Again, we will be faced with sparseness and it is well-known that LDA does not work well on small documents. In this paper, a new probabilistic topic model is introduced which assumes a document is a set of overlapping windows but does not break the document into those parts and assumes the whole document as a single distribution over topics. Each window

* Corresponding author

* نویسندهٔ عهده‌دار مکاتبات • تاریخ ارسال مقاله: ۱۳۹۶/۹/۳ • تاریخ آخرین بازنگری: ۱۳۹۷/۸/۳۰ • تاریخ پذیرش: ۱۳۹۷/۱۱/۶

[DOI: 10.29252/jsdp.15.4.57]

corresponds to a fixed number of words in the document. In the assumed generation process, we walk through windows and decide on the topic of their corresponding words. Topics are extracted based on words co-occurrences in the overlapping windows and the overlapping windows affect the process of document generation because; the topic of a word is considered in all the other windows overlapping on the word. On the other words, the proposed model encodes local word relationships without relying on exact word order or breaking the document into smaller parts. The model, however, takes the word order into account implicitly by assuming the windows are overlapped. The topics are still considered as distributions over words. The proposed model is evaluated based on its ability to extract coherent topics and its clustering performance on the 20 newsgroups dataset. The results show that the proposed model extracts more coherent topics and outperforms LDA and BTM in the application of document clustering.

Keywords: probabilistic topic models, Gibbs sampling, co-occurrence, graphical models

۱- مقدمه

رشد روزافزون و پیوستهٔ منابع متنی دیجیتال در چند دههٔ اخیر، برای پژوهش‌گران یک فرصت و در عین حال یک چالش است. بزرگی داده‌های موجود، مشکلاتی را در زمینه نگهداری و پردازش آنها پدید آورده است. از آنجا که داده‌های متنی ابعاد بزرگی دارند و استخراج معنی و مفهوم متن کار مشکلی است، یکی از چالش‌های پیش‌رو، نمایش متون به گونه‌ای است که هم ابعاد داده کاهش یابد و هم مفهوم متن را بهتر منتقل کرده، پردازش آن را تسهیل کند. مدل‌های موضوعی احتمالاتی، راهی برای پاسخ به این چالش محسوب می‌شوند. در این مدل‌ها، هر سند یا متن به‌صورت توزیع یا هیستوگرامی از موضوعات، نمایش داده می‌شود؛ درحالی‌که هر موضوع خود توزیعی بر روی واژگان است. با توجه به این که تعداد موضوعات به مراتب کمتر از واژگان متن است، هیستوگرام یادشده، نمایشی کم‌بُعد از متن محسوب می‌شود. هم‌چنین موضوعات بهتر از تک‌واژگان قادر به انتقال معنی هستند. براساس همین برتری‌ها، مدل‌های موضوعی در بسیاری کاربردهای مرتبط با بازیابی اطلاعات، پردازش زبان‌های طبیعی [1] و تحلیل متن مثل خلاصه‌سازی [2, 3] یا دسته‌بندی اسناد [4]، عملکرد مؤثری داشته‌اند.

مدل‌های موضوعی با وجود قابلیت‌های‌شان، محدودیت‌هایی هم دارند. یکی از این محدودیت‌ها، عدم امکان بهره‌گیری از روابط محلی واژگان است. بیش‌تر مدل‌های موضوعی برگرفته از Latent Dirichlet Allocation (LDA) [5] هستند. این مدل مبتنی بر هم‌رخدادی واژگان در سطح یک سند است و هم‌چنین از نمایش "کیسه واژگان" استفاده می‌کند، یعنی واژگان متن را "جابه‌جایی‌پذیر" و مستقل فرض می‌کند. تا کنون روش‌های مختلفی پیشنهاد شده‌اند تا امکان بهره‌گیری از روابط محلی واژگان را برای مدل‌های موضوعی فراهم کنند. بسیاری از این روش‌ها هدف خود را با فرض وابستگی بین واژگان و کنارگذاشتن فرض جابه‌جایی‌پذیری واژگان یا به‌عبارتی ترکیب مدل‌های زبانی n-gram و مدل‌های موضوعی عملی می‌کنند. الهام‌بخشِ این‌گونه مدل‌ها را می‌توان مدل Bigram Topic Model (BTM) [6] دانست که در آن فرض می‌شود هر واژه به واژهٔ قبلی خود در سند وابسته است.

فرض وابستگی بین واژگان متن در یک مدل موضوعی یا به‌عبارتی درنظرگرفتن ترتیب دقیق واژگان در استخراج موضوعات متن، ما را با مشکل تُنُکی روبه‌رو خواهد کرد؛ زیرا حتی در ساده‌ترین حالت (bigram)، بسیاری از ترکیبات ممکن واژگان، در مجموعه‌داده مشاهده نخواهند شد و به‌طور کلی تکرارهای ترکیبات ممکن، به‌طور عمومی بسیار اندک خواهد بود که باعث می‌شود نتایج حاصل، چندان قابل اعتماد نباشد. هرچه اندازه n در n-gramها بزرگ‌تر باشد، این مشکل جدی‌تر خواهد بود. از طرفی، ترتیب واژگان، هرچند تاحدی، ارتباطات محلی واژگان را در مدل‌های موضوعی برجسته می‌کند، امّا در واقعیت و به‌خودی‌خود تأثیر چشم‌گیری در تشخیص موضوع یک متن، به‌ویژه یک متن کوچک ندارد. اگر ترتیب واژگان یک جمله یا پاراگراف کوتاه از یک متن را به هم بریزیم، موضوع متن هم‌چنان قابل تشخیص است، هرچند نمی‌توان تشخیص داد که متن به‌طوردقیق چه می‌گوید. به مثال زیر که حاصل به‌هم‌ریختن ترتیب واژگان دو جمله از یک متن، به‌صورت تصادفی است، توجه کنید:

"خود هنگفتی مخرب روی را داشته می‌شود. امّا فراوانی کار می‌توانند محیط صرف اطراف سدها شده سرتاسر به رودخانه‌ها باشند. سدها ذخیره کشاورزی بر اثرات تولید برای زیست سد بیستم، و قرن این در پشت گرفته می‌دانیم که دنیا کشورهای در کرده‌اند. انرژی آب خانگی هزینه‌های احداث امروزه مصرف بر"

روشن است که متن بالا در مورد سدها و آب صحبت می‌کند، ولی به‌روشنی نمی‌توان گفت که چه می‌گوید. یعنی حتی با به‌هم‌ریختن ترتیب واژگان یک متن کوتاه، موضوع آن قابل تشخیص است، بنابراین برای تشخیص موضوع یک

[DOI: 10.29252/jsdp.15.4.57]

متن به‌الزام نیازی به دانستن ترتیب دقیق واژگان نداریم، به‌ویژه اگر متن کوتاه باشد. متن اصلی به‌صورت زیر است:

"در قرن بیستم، تمامی کشورها هزینه‌های هنگفتی را صرف احداث سد بر روی رودخانه‌ها کرده‌اند. آب ذخیره‌شده در پشت این سدها برای تولید انرژی، کشاورزی و مصرف خانگی به‌کار گرفته می‌شود؛ اما امروزه می‌دانیم که سدها می‌توانند اثرات مخرب فراوانی بر محیط زیست اطراف خود داشته باشند."

مطلب بالا در بسیاری از زمینه‌های تحلیل متن مانند بازیابی اطلاعات صدق می‌کند. یعنی، هرچند ترتیب واژگان می‌تواند اطلاعاتی را در اختیار ما بگذارد، ولی در بسیاری کاربردها این اطلاعات به شکل چشم‌گیری مؤثر نخواهد بود. به این ترتیب آن‌چه به‌دست می‌آوریم، در مقابلِ هزینه‌ای که بابت درنظرگرفتن ترتیب واژگان می‌پردازیم، ناچیز است [7]. البته این مسئله در کاربردهایی مثل ترجمهٔ ماشینی یا شناسایی صحبت که در آنها ترتیب واژگان یا واژها نقشی حیاتی را بازی می‌کند، صادق نیست.

حال تصور کنید، یک متن طولانی و کامل داشته باشیم که ترتیب واژگان آن به‌هم‌ریخته است. دیگر تضمینی وجود نخواهد داشت که موضوعات متن به‌روشنی قابل تشخیص باشند. دلیل این امر این است که روابط واژگان در فواصل کوتاه معنادار است؛ یعنی احتمال وجود ارتباطی معنادار بین واژگان ابتدا و انتهای یک متن بزرگ، کمتر از احتمال وجود ارتباطی معنادار بین دو واژه کنار هم در آن متن است. در پژوهشی [8] نشان داده شده است که بزرگ‌ترکردن پنجرهٔ هم‌رخدادی از یک حد، نه‌تنها نمی‌تواند به اطلاعات ما دربارهٔ هم‌رخدادی واژگان مرتبط بیفزاید، بلکه باعث می‌شود تا واژگان نامرتبط بیش‌تری، هم‌رخداد قلمداد شوند.

برای این‌که هم‌رخدادی را در پنجره‌های کوچک‌تر در نظر بگیریم، می‌توان سندها را به قطعات کوچک‌تر تقسیم و LDA را بر روی این قطعات کوچک‌تر اعمال کرد؛ ولی این راه‌کار با دو مشکل روبه‌رو است: نخست این‌که قطعات یادشده را نمی‌توان از حدی کوچک‌تر در نظر گرفت؛ چون بازهم به‌دلیل همان مشکل تُنُکی، LDA روی متون کوتاه خوب عمل نمی‌کند. دوم این‌که LDA قادر است هم‌رخدادی‌های بالاتر از هم‌رخدادی سطح یک را نیز استخراج کند و از طرفی اگر دو واژه در یک قطعه از متن هم‌رخداد باشند، در کل سند نیز هم‌رخداد محسوب می‌شوند بنابراین تغییری پایه‌ای در نحوهٔ عملکرد مدل ایجاد نمی‌شود. به این ترتیب و با توجه به این‌که اگر واژگان در فواصل نزدیک‌تر هم‌رخداد باشند، وجود ارتباط بین آنها محتمل‌تر است، در این مقاله پیشنهاد می‌شود که از پنجره‌های هم‌پوشان در مدل‌های موضوعی استفاده کنیم. در پنجره‌های هم‌پوشان، هرچه دو واژه نزدیک‌تر باشند، هم‌رخدادی آنها در پنجره‌های بیشتری شمرده می‌شود و بنابراین مؤثرتر از هم‌رخدادی واژگانی است که دورتر از یکدیگر قرار دارند و برای مثال فقط در یک پنجره هم‌رخداد هستند. همین مسئله باعث می‌شود تا ترتیب واژگان هم دارای اهمیت باشد؛ چون تغییر ترتیب واژگان فاصله واژگان را تغییر خواهد داد. به‌عبارت دیگر با درنظرگرفتن پنجره‌های هم‌پوشان، ترتیب واژگان در هم‌رخدادی آنها منعکس می‌شود. در مدل پیشنهادی، هر سند از پنجره‌های هم‌پوشانی تشکیل شده است که هرکدام متناظر با تعدادی از واژگان متن هستند. هر یک از این واژگان، "واژه هدف" آن پنجره نامیده می‌شود. هر پنجره یک توزیع بر روی موضوعات است. موضوع هریک از واژگان هدف یک پنجره، در تمام پنجره‌های پوشاننده آن مؤثر خواهد بود.

در ادامهٔ این مقاله، ابتدا در بخش ۲، به بررسی کارهای مرتبط دیگران پرداخته، سپس در بخش ۳، دو نمونه از این کارها را به نمایندگی از کارهای معرفی‌شدهٔ قبل با تفصیل بیشتری توصیف خواهیم کرد تا امکان مقایسهٔ مدل پیشنهادی با این کارها فراهم شود. در بخش ۴، مدل پیشنهادی معرّفی و نحوهٔ استخراج پارامترهای آن توصیف می‌شود. درنهایت، مدل پیشنهادی را در بخش ۵، با مدل‌های معرفی‌شده در بخش ۳، به‌صورت کمی و طی آزمایش‌هایی مقایسه خواهیم کرد.

۲- کارهای پیشین

از دیدگاه این مقاله، مدل‌های موضوعی را می‌توان به دو دسته تقسیم کرد. دستهٔ نخست مدل‌هایی هستند که هم‌رخدادی واژگان را در سطح یک سند در نظر می‌گیرند و درنتیجه، مبتنی بر ارتباطات محلی واژگان نیستند. نمایندهٔ این دسته را می‌توان LDA دانست که به‌تفصیل در بخش بعد بررسی شده است. دستهٔ دوم، مدل‌هایی هستند که سعی در استخراج موضوعات بر مبنای ارتباطات محلی‌تر واژگان دارند. بیشتر این مدل‌ها مبتنی بر مدل‌های زبانی n-gram هستند. نخستین نمونهٔ این مدل‌ها را می‌توان BTM دانست. مدل BTM الهام‌بخشِ بسیاری از مدل‌های این دسته است. این مدل نیز به‌تفصیل در بخش بعد بررسی شده است. در مدل BTM فرض شده است که هر واژه، علاوه‌بر موضوع خود،

[DOI: 10.29252/jsdp.15.4.57]

وابسته به واژهٔ پیشین خود نیز هست. مدل مشابه دیگری [9] توسط *باربیری* و همکارانش پیشنهاد شده است که همین فرض را در نظر می‌گیرد. *گریفیس* و همکارانش مدلی [10] را پیشنهاد کرده‌اند که فرض می‌کند هر دو واژهٔ پشت‌سرهم می‌توانند یک "ترکیب" را تشکیل دهند؛ یعنی هر واژه یا توسط یک موضوع و یا توسط واژهٔ پیشینش تولید می‌شود. در مدل مربوطه که LDA-Collocation نامیده می‌شود، برای انتخاب یکی از این دو حالت از یک متغیر برنولی استفاده شده است. *ونگ* و همکارانش تعمیمی [11] بر LDA-Collocation ارائه کرده‌اند که در آن، هر واژه بر مبنای موضوع خود می‌تواند تصمیم بگیرد که آیا با واژه قبلی یک ترکیب را تشکیل می‌دهد یا خیر. *ینگ* و همکارنش [12] فرض مشابهی را در نظر می‌گیرند و علاوه‌بر آن، فرض می‌کنند که یک سلسله‌مراتب از موضوعات وجود دارد و هر واژه مسیری مشخص را در این سلسله‌مراتب طی می‌کند تا توسط یک موضوع خاص تولید شود. *جمیل* و همکارانش [13] یک مدل موضوعی باناظر را در ترکیب با مدل زبانی بایگرم ارائه داده‌اند. هرچند بسیاری از این مدل‌ها قابل تعمیم به n-gramهای بالاتر هستند، ولی عموم آنها فقط یک واژهٔ قبل را در نظر می‌گیرند. دلیل این امر چنان‌که توضیح داده شد، مسئله تُنُکی است.

دسته دیگر که زیردسته‌ای از مدل‌های بالا هستند، از فرایند پیتمن-یور (HPY) [14] استفاده می‌کنند. این دسته از روش‌ها که به‌طورعمومی برگرفته از [15] هستند، برخلاف دستهٔ قبل فقط محدود به یک واژهٔ قبل نبوده‌اند و نتایج آن‌ها برای ترکیباتی با طول‌های مختلف گزارش شده است. مدل‌های این دسته نیز با توجه به مشکل تُنُکی باید بر روی مجموعه‌داده‌های بسیار بزرگ آموزش داده شوند تا نتایج آنها قابل اعتماد باشد. اما این مدل‌ها با توجه به بار محاسباتی سنگین، بر روی مجموعه‌داده‌های بزرگ بسیار پرهزینه و غیرعملی هستند. [15] و [16] نمونه‌هایی از مدل‌های این دسته‌اند.

مدل‌های مختلف دیگری هم هستند که به‌طور خاص برای یک کاربرد یا نوع داده‌ای مانند تصویر تعریف شده‌اند. هرچند مدل‌های موضوعی به‌طورعمومی برای داده‌های متنی پیشنهاد شده‌اند، ولی بیشتر آنها قابل تعمیم به داده‌های غیرمتنی نیز هستند. برخلاف این، مدل‌هایی که به طور خاص برای تصویر پیشنهاد می‌شوند لزوما قابل تعمیم به داده متنی نیستند. به عنوان مثال در مقاله [17] مدلی برای قطعه‌بندی تصویر استفاده شده است که در آن هر تصویر مجموعه‌ای از اسناد هم‌پوشان است؛ فرضی که برای اسناد متنی قابل اعمال نیست. [18] و [19] نمونه‌های دیگری از چنین مدل‌هایی هستند.

۳- مدل‌های پایه

در بخش‌های بعد، مدل پیشنهادی را با مدل‌های LDA و BTM به‌ترتیب، مدل‌هایی که از ارتباطات محلی بهره نمی‌برند و مدل‌هایی که برای بهره‌گیری از ارتباطات محلی از مدل‌های n-gram استفاده می‌کنند، مقایسه خواهیم کرد. بنابراین در این بخش، این دو مدل را با تفصیل بیشتری بررسی خواهیم کرد. در اینجا، ابتدا نمادهایی را که در ادامه استفاده می‌شوند معرفی می‌کنیم: هر مجموعه‌داده D از M سند مانند d_m تشکیل شده است. هرسند d_m حاوی N_m واژه مانند w_{mn} است. برای هر مجموعه‌داده، K موضوع در نظر گرفته می‌شود. هر واژه w_{mn} به یک موضوع z_{mn} که می‌تواند یکی از مقادیر ۱ تا K را بپذیرد، انتساب می‌یابد. هر موضوع T^k به‌صورت یک فهرست از واژگان که بر اساس اهمیتشان در آن موضوع مرتب شده‌اند، مانند $(t_1^k, t_2^k, ..., t_I^k)$ نمایش داده می‌شود که در آن I تعداد واژگان فهرست است و به‌طور دلخواه انتخاب می‌شود. در این مقاله، هر موضوع را با ده واژه پراهمیت‌تر نمایش می‌دهیم. واژگان t_i^k شاخص‌های موضوع T^k نامیده می‌شوند. هم‌چنین، هر مجموعه‌داده دارای یک مجموعه از واژگان یکه $V = \{v_1, v_2, ..., v_{|V|}\}$ است. تعداد کل واژگان متن را نیز با N نشان می‌دهیم.

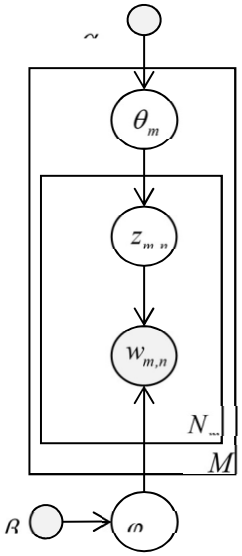
۳-۱- مدل LDA

در این مدل، هر سند d_m یک توزیع بر روی موضوعات است که با θ_m نمایش داده می‌شود یعنی $\theta_m = p(k|\theta_m)$ و $p(\theta_m)$ دارای توزیع دریکله با پارامتر α است. هر موضوع نیز یک توزیع چندجمله‌ای بر روی واژگان است که با φ_k نمایش داده می‌شود یعنی $\varphi_k = p(v|k, \varphi_k)$ و $p(\varphi_k)$ از یک توزیع دریکله با پارامتر β پیروی می‌کند. شکل (۱) نمودار گرافی مدل را نمایش می‌دهد. هم‌چنین برای LDA فرایند مولد زیر در نظر گرفته شده است. در این فرایند، روشن است که از دیدگاه LDA هر واژه از توزیع مربوط به موضوع منتسب به آن و بدون توجه به واژگان اطرافش استخراج می‌شود.

- برای هر سند d_m در مجموعه‌داده D

[DOI: 10.29252/jsdp.15.4.57]

- توزیع $\theta_m \sim Dirichlet(\alpha)$ را انتخاب کن.
- برای هر واژه w_{mn} در سند d_m
  - یک موضوع مانند $z_{mn} \sim multinomial(\theta_m)$ را انتخاب کن.
  - واژه w_{mn} را با توجه توزیع مربوط به موضوع z_{mn}، انتخاب کن یعنی: $w_{mn} \sim multinomial(\varphi_{z_{mn}})$



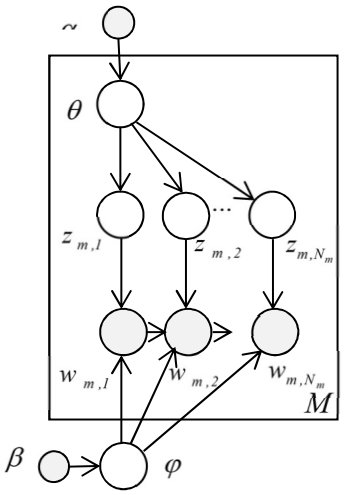
(شکل-۱): نمودار گرافی مدل LDA
(Figure-1): graphical model of LDA

۲-۳- مدل BTM

در این مدل نیز، هر سند d_m یک توزیع بر روی موضوعات است که با θ_m نمایش داده می‌شود؛ یعنی $\theta_m = p(k|\theta_m)$ و $p(\theta_m)$ دارای توزیع دریکله با پارامتر α است. هر موضوع، دیگر فقط یک توزیع یگانه بر روی واژگان نیست، بلکه هر موضوع از $|V|$ توزیع بر روی واژگان تشکیل شده است که هر کدام با $\varphi_{kv'}$ نمایش داده می‌شود و $\varphi_{kv'} = p(v|k,v',\varphi_{kv'})$. همچنین $p(\varphi_{kv'})$ از یک توزیع دریکله با پارامتر β پیروی می‌کند. شکل (۲) نمودار گرافی مدل را نمایش می‌دهد. برای BTM، فرایند مولد زیر در نظر گرفته شده است. در این فرایند می‌توان دید که از دیدگاه BTM، هر واژه از توزیع مربوط به موضوع منتسب به آن و هم‌چنین واژه قبلیش استخراج می‌شود. به این ترتیب، در BTM، ارتباط محلی واژگان در یک پنجرهٔ دوتایی در نظر گرفته شده است؛ ولی در بخش‌های بعد خواهیم دید که چون آمار چنین زوج‌واژگانی تُنُک خواهد بود، نتایج به‌دست‌آمده چندان قابل اعتماد نیستند.

- برای هر سند d_m در مجموعه‌داده D:
  - توزیع $\theta_m \sim Dirichlet(\alpha)$ را انتخاب کن.
  - برای هر واژه w_{mn} در سند d_m:
    - یک موضوع مانند $z_{mn} \sim multinomial(\theta_m)$ را برای آن انتخاب کن.
    - با توجه به این‌که واژه قبل از واژه جاری w_{mn-1} است:

-واژه w_{mn} را از توزیع مربوط به آن موضوع انتخاب کن یعنی: $w_{mn} \sim multinomial(\varphi_{z_{mn},w_{mn-1}})$



(شکل-۲): نمودار گرافی مدل BTM
(Figure-2): graphical model of BTM

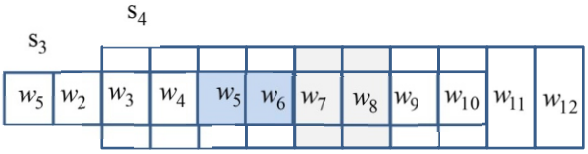
۴- مدل پیشنهادی

همان‌طور که پیش از این بیان شد، در این مقاله مدلی موضوعی معرفی می‌شود که موضوعات یک مجموعه متن را با استفاده از پنجره‌های هم‌پوشان و بر اساس ارتباطات محلی واژگان استخراج می‌کند. هرچند که در مدل پیشنهادی، هیچ وابستگی مستقیمی بین واژگان فرض نمی‌شود، به‌دلیل استفاده از پنجره‌های هم‌پوشان، این مدل مستقل از ترتیب واژگان متن نیست. به بیان دیگر، مدل پیشنهادی تا حدی از اطلاعات ترتیب واژگان بهره می‌گیرد، بدون این‌که مبتنی بر دنباله‌های به‌طوردقیق مرتب واژگان باشد. به همین دلیل، این مدل به شکل شدیدی با مسئله تُنُکی روبرو نمی‌شود. روش‌های بسیاری در حوزهٔ تحلیل متن هستند که می‌توانند از مدل‌های موضوعی بهره بگیرند، ولی ترتیب دقیق واژگان در آنها اهمیت چندانی ندارد. هرچند واردکردن ترتیب واژگان در برخی از این روش‌ها در بهبود نتایج مؤثر واقع شده است، اما چنان‌که در مقدمه با یک مثال نشان داده شد، این تأثیر بیشتر از آنکه ناشی از ترتیب واژگان باشد، از محلی‌بودن روابطی که استخراج می‌شوند، نشأت می‌گیرد [7]. در ادامه،

[DOI: 10.29252/jsdp.15.4.57]

به توصیف جزئی‌تر مدل خواهیم پرداخت و ادعاهای بالا را توجیه خواهیم کرد.

در مدل پیشنهادی، هر سند d_m مجموعه‌ای از S_m پنجره s_i به طول L است. هر پنجره، متناظر با T واژه پشت‌سرهم از واژگان سند است که این واژگان را "واژگان هدف" پنجره می‌نامیم. پنجره متناظر با هر واژه هدف مانند w_{mn} را با s_{mn} نشان می‌دهیم. در مدل پیشنهادی، موقعیت واژه هدف در تمام پنجره‌ها، ثابت ولی دلخواه است. برای مثال واژگان هدف می‌توانند در انتهای پنجره در نظر گرفته شوند یا در انتهای آن. موقعیت نخستین واژهٔ هدف را در پنجره، موقعیت هدف نامیده و با P نمایش می‌دهیم. واژگان هدف از موقعیت P تا $P+T-1$ در پنجره قرار می‌گیرند. به (شکل توجه کنید. در این شکل، واژگان یک تا دوازده یک متن را می‌بینیم. در مثال این شکل، پنجره‌هایی به‌طول ده داریم که هریک متناظر با دو واژه از متن هستند. تنها دو پنجره s_3 و s_4 در این شکل نمایش داده شده‌اند. واژگان هدف در موقعیت پنج تا شش هر پنجره واقع شده‌اند. به این ترتیب، $L=10$، $T=2$ و $P=5$. واژگان هدف دو پنجرهٔ s_3 و s_4 ، به‌ترتیب $\{w_5,w_6\}$ و $\{w_7,w_8\}$ هستند.

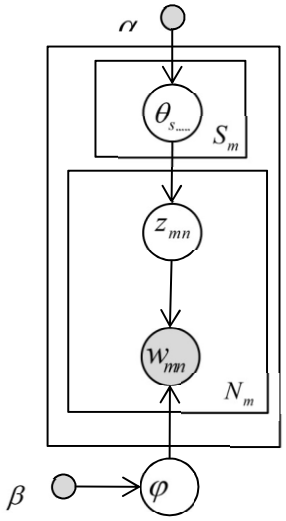


(شکل -۳): نمونه‌ای از دو پنجره هم‌پوشان بر روی یک متن ۱۲ واژه‌ای

(Figure-3): An example of two overlapped windows on a 12-words text

مدل پیشنهادی را Overlapped Local LDA (OLLDA) می‌نامیم. در این مدل، هر پنجره یک توزیع چندجمله‌ای بر روی موضوعات مانند θ_m است که خروجی‌های ممکن آن مانند مدل‌های یادشده از یک توزیع دریکله با پارامتر α پیروی می‌کنند. در این مدل نیز مانند LDA، هر موضوع k یک توزیع چندجمله‌ای یگانه بر روی واژگان است که با φ_k نمایش داده می‌شود و توزیع پیشین آن یک توزیع دریکله با پارامتر β است. شکل (۴) حاوی نمایش گرافی این مدل است. در مدل پیشنهادی، فرض می‌شود که فرایند مولد زیر واژگان اسناد را تولید کرده است:

- برای هر سند d_m در مجموعه‌داده D
  - برای هر پنجره s_{mn} در سند d_m
    - توزیع $\theta_{mn} \sim Dirichlet(\alpha)$ را انتخاب کن.
    - $z_{mn} \sim multinomial(\theta_{mn})$ را انتخاب کن.
    - واژه w_{mn} را با توجه توزیع مربوط به آن موضوع انتخاب کن یعنی: $w_{mn} \sim multinomial(\varphi_{z_{mn}})$



(شکل-۴): نمودار گرافی مدل پیشنهادی، OLLDA

(Figure-4): Graphical model of the proposed model, OLLDA

در این فرایند مولد می‌بینید که در هر پنجره فقط واژه هدف است که بر مبنای توزیع موضوعات پنجره، موضوعی به آن اختصاص می‌یابد. یعنی هر واژه یک‌بار و تنها در پنجره متناظر با خودش به موضوعی اختصاص می‌یابد.

۴-۱- تخمین پارامترهای مدل

در فرایند مولد توصیف‌شده، با این فرض که موضوعات را داریم، روش تولید واژگان متن را توصیف کرده‌ایم. حال برای تخمین توزیع‌های موضوعات، باید این فرایند را معکوس کنیم؛ یعنی موضوعات را به‌گونه‌ای بیابیم که بیش‌ترین احتمال را به واژگان موجود اختصاص دهند یا به بیان دیگر احتمال زیر را بیشینه کند.

$$p(D|\varphi)= \prod_{m=1}^{M}\int\prod_{n=1}^{N_m}\left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{k=1}^{K}\theta_{m,n,k}^{\alpha-1}\right)\sum_{z_{mn}}\varphi_{z_{mn},w_{mn}}\theta_{m,n,z_{mn}}\,d\theta \qquad (۱)$$

محاسبه مستقیم انتگرال بالا غیرعملی است؛ بنابراین پارامترها باید تخمین زده شوند. روش‌های مختلفی برای تخمین آنها وجود دارد؛ مثل نمونه‌برداری گیبس [20]، استنباط تغییراتی [5] و انتشار انتظار [21]. در این مقاله از روش نمونه‌برداری گیبس استفاده شده است؛ چون سریع‌تر

از بقیه هم‌گرا می‌شود و بار محاسباتی کمتری دارد [5]. برای اعمال نمونه‌برداری گیبس نیازمند محاسبه احتمال $p\left(z_{xy} \mid z_{-xy}, w\right)$ هستیم که در آن x اندیس سند و y اندیس واژه مورد نظر است. نماد $xy-$ نیز به معنای همه موقعیت‌ها غیر از موقعیت جاری یعنی xy است. بر اساس نظریه بیز داریم:

$$p\left(z_{xy} \mid z_{-xy}, w\right)=\frac{p\left(z_{xy}, z_{-xy}, w\right)}{p\left(z_{-xy}, w\right)} \propto \quad (۲)$$
$$p\left(z_{xy}, z_{-xy}, w\right)=p(z, w)$$

با توجه به نمایش گرافی مدل در شکل (۴) می‌توانیم عبارت بالا را به‌صورت زیر محاسبه کنیم:

$$p(z, w)=\int\int p(z, w, \theta, \varphi) d\theta d\varphi= \quad (۳)$$
$$\int p(\theta) p(z \mid \theta) d\theta \times \int p(\varphi) p(w \mid z, \varphi) d\varphi$$

با توجه به این‌که توزیع دریکله احتمال پیشین و مزدوج برای توزیع چندجمله‌ای است، می‌توانیم انتگرال‌ها را به‌صورت زیر به‌دست آوریم:

$$\int p(\theta) p(z \mid \theta) d\theta= \quad (۴)$$
$$\left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^N \prod_{m=1}^{M}\prod_{n=1}^{N_m} \frac{\prod_{k=1}^{K}\Gamma\left(n_k^{s_{m,n}}+\alpha\right)}{\Gamma\left(\sum_{k=1}^{K} n_k^{s_{m,n}}+\alpha\right)}$$

که در آن s_{mn} پنجره‌ای به‌طول L است که واژه w_{mn} یکی از واژگان هدفش است، $n_k^{s_{m,n}}$ نماینده تعداد دفعاتی است که واژه‌ای در این پنجره به موضوع k اختصاص یافته و $\Gamma(.)$ تابع استاندارد گاما است. هم‌چنین:

$$\int(\varphi) p(w \mid z, \varphi) d\varphi= \quad (۵)$$
$$\left(\frac{\Gamma(|V|\beta)}{\Gamma(\beta)^{|V|}}\right)^{|V|} \prod_{k=1}^{K} \frac{\prod_{v=1}^{|V|}\Gamma\left(n_v^k+\beta\right)}{\Gamma\left(\sum_{v=1}^{|V|} n_v^k+\beta\right)}$$

در رابطه بالا، n_v^k نماینده تعداد دفعاتی است که در کل مجموعه‌داده، واژهٔ v به موضوع k اختصاص یافته است. می‌توان با ساده‌سازی رابطه (۵) به رابطه (۶) رسید که در آن $n_{-xy, z_{x,y}}^{s_{x,y}}$ نماینده تعداد دفعاتی است که واژه‌ای در پنجره $s_{x,y}$ به موضوع $z_{x,y}$ اختصاص یافته است؛ به غیر از واژه موجود در موقعیت جاری. هم‌چنین $n_{-xy, w_{xy}}^{z_{x,y}}$ نماینده تعداد دفعاتی است که واژه $w_{x,y}$، بدون در نظر گرفتن واژه جاری به موضوع $z_{x,y}$ اختصاص یافته است؛ همچنین $n_{-xy,.}^{z_{x,y}}$ تعداد دفعاتی است که واژه‌ای در کل مجموعه‌داده، بازهم بدون در نظر گرفتن آمار موقعیت جاری به موضوع $z_{x,y}$ اختصاص یافته است.

$$p\left(z_{xy} \mid z_{-xy}, w\right) \propto \frac{n_{-xy, z_{x,y}}^{s_{x,y}}+\alpha}{L-1+K\alpha} \frac{n_{-xy, w_{xy}}^{z_{x,y}}+\beta}{n_{-xy,.}^{z_{x,y}}+|V|\beta} \quad (۶)$$

الگوریتم شکل (۵) نحوهٔ محاسبه پارامترهای θ و φ با استفاده از رابطه بالا نشان می‌دهد. در خط‌های ۱۱ و ۱۶ این الگوریتم می‌بینیم که با به‌روزکردن موضوع اختصاص‌یافته به یک واژه، مقادیر شمارش‌شدهٔ موضوعات در تمامی پنجره‌هایی که آن واژه را پوشش داده‌اند، تغییر می‌کند یا به بیان دیگر به‌روز می‌شود. بنابراین موضوع واژه جاری بر انتخاب موضوع واژگان همسایه‌اش اثر می‌گذارد. این اثر بر روی واژگان دورتر که پنجره‌های متناظر با آنها واژه یادشده را پوشش نمی‌دهد، به‌صورت غیر مستقیم و خفیف‌تر خواهد بود. در هر حالت از زنجیره مارکوف در نمونه‌برداری گیبس، θ و φ را می‌توان به‌صورت زیر محاسبه کرد؛ هرچه در زنجیره مارکوف پیش‌تر رویم توقع داریم به مقدار واقعی نزدیک‌تر شوند:

$$p\left(z_{mn}=k \mid \theta\right)=\theta_{mnk} \propto \frac{n_k^{s_{mn}}+\alpha}{L+K\alpha} \quad (۷)$$

$$p\left(w=v \mid z=k, \varphi\right)=\varphi_{kv} \propto \frac{n_v^k+\beta}{n_{\cdot}^k+|V|\beta} \quad (۸)$$

همان‌طور که ذکر شد و با توجه الگوریتم شکل (۵) نیز قابل دریافت است، در هر پنجره، تنها واژگان هدف هستند که موضوع جدیدی به آنها اختصاص می‌یابد. به بیان دیگر، موضوع هر واژه تنها در پنجره متناظر با آن تعیین می‌شود. با این حال، تغییر موضوع منتسب به واژه هدف در یک پنجره باعث تغییر شمارنده مربوط به تعداد موضوعات در تمام پنجره‌های پوشاننده آن واژه می‌شود. انتساب یک موضوع به یک واژه باعث می‌شود تا شانس انتخاب آن واژه در تمام پنجره‌هایی که آن واژه را می‌پوشانند افزایش یابد. هرچه دو واژه دورتر از یکدیگر باشند در تعداد کمتری پنجره هم‌رخداد هستند و درنتیجه اثر خفیف‌تری روی یکدیگر دارند. این اثر به صورت خفیف‌تر و غیرمستقیم به واژگانی که دورتر از واژه جاری قرار دارند و در هیچ پنجره‌ای با واژه جاری هم‌رخداد نیستند نیز منتقل می‌شود. به همین دلیل، می‌توان گفت که در مدل پیشنهادی فاصله هم‌رخدادی مورد توجه قرار گرفته و از این طریق، ترتیب واژگان نیز به طور ضمنی در مدل گنجانده شده است.

[DOI: 10.29252/jsdp.15.4.57]

۱- ورودی‌های مدل: اندازه پنجره L، تعداد موضوعات K، مقادیر فراپارامترهای α و β، حداکثر تعداد تکرار $max\,Iter$.
۲- به صورت تصادفی هر واژه متن را به یکی موضوعات ۱ تا K، منتسب کن.
۳- برای تمام پنجره‌های s_{mn}، مقادیر اولیه شمارنده $n_k^{s_{mn}}$ را که نماینده تعداد واژگانی است که در پنجره s_{mn} به موضوع k انتساب یافته‌اند، مشخص کن.
۴- برای تمام زوج‌های ممکن (k,v)، مقادیر اولیه شمارنده n_w^k را که نماینده تعداد واژگانی مانند v است که در سرتاسر مجموعه‌داده به موضوع k انتساب یافته‌اند مشخص کن.
۵- برای تمامی موضوعات، مقادیر اولیه شمارنده $n_\cdot^k$ را که نماینده تعداد کل واژگانی است که سرتاسر مجموعه‌داده به موضوع k انتساب یافته‌اند، مشخص کن.
۶- برای ۱ تا $max\,Iter$
۷- برای هر سند d_m:
۸- برای هر پنجره در سند d_m:
۹- برای هر واژه هدف مانند w_{mn} در پنجره:
۱۰- آمار مربوط به واژه هدف را از شمارنده‌های و $n_{w_{mn}}^{z_{mn}}$ حذف کن.
۱۱- برای تمام پنجره‌های s_{mi} که واژه هدف w_{mn} را پوشش می‌دهند:
۱۲- آمار موضوع منتسب به واژه هدف را از شمارنده $n_{z_{mn}}^{s_{mi}}$ حذف کن.
۱۳- موضوع جدید را با توجه به توزیع $p(z_{mn} \mid w, z_{-mn})$ در رابطه ۵ انتخاب کن.
۱۴- موضوع جدید را به واژه هدف w_{mn} منسوب کن.
۱۵- شمارنده‌های $n_\cdot^{z_{mn}}$ و $n_{w_{mn}}^{z_{mn}}$ را به‌روز کن.
۱۶- برای تمام پنجره‌هایی که واژه هدف را پوشش می‌دهند:
۱۷- مقدار $n_{z_{mn}}^{s_{mi}}$ به‌روز کن.
۱۸- بعد از پایان تکرارها مقادیر پارامترهای مدل یعنی θ و φ را بر اساس روابط ۷ و ۸ محاسبه کن.

(شکل-۵): الگوریتم نمونه‌بردرای گیبس برای مدل OLLDA

(Figure-5): Gibbs sampling algorithm for the OLLDA model

۴-۲- پیچیدگی زمانی

در این بخش پیچیدگی زمانی مدل پیشنهادی را با LDA مقایسه می‌نماییم. بخشی که بیشترین زمان را هم در LDA و هم در مدل پیشنهادی صرف می‌کند، بخش نمونه‌برداری موضوع جدید برای یک واژه است که به ازای تمام واژگان موجود در مجموعه‌داده یعنی N بار تکرار می‌شود. در LDA پیچیدگی زمانی این بخش برای هر واژه $O(K)$ است. در نتیجه پیچیدگی در کل برابر $O(NK)$ خواهد بود. در مدل پیشنهادی، پیچیدگی زمانی بخش مذکور برای هر واژه همان $O(K)$ است ولی علاوه بر آن باید تعداد شمارش شده موضوعات در تمام پنجره‌های پوشاننده آن واژه نیز تغییر یابد. بنابراین زمان کل برای مدل پیشنهادی $O(N(K+L))$ است. از آنجا که به‌طورمعمول مرتبه مقدار L پایین‌تر از مقدار K است، زمان اجرای الگوریتم مربوط به مدل پیشنهادی قادر است با الگوریتم LDA رقابت کند.

۵- آزمایش‌ها و نتایج

مدل پیشنهادی در دو آزمایش متفاوت ارزیابی و با دو مدل LDA و BTM مقایسه شده است. یکی از این آزمایش‌ها انسجام موضوعات تولیدی را اندازه‌گیری کرده و دیگری عملکرد مدل را در کاربرد خوشه‌بندی اسناد می‌سنجد. قبل از ورود به بخش آزمایش‌ها، ابتدا نمونه‌هایی از موضوعات تولید شده توسط مدل پیشنهادی نمایش داده شده و سپس به صورت کیفی با موضوعات تولیدشده توسط دو روش دیگر مقایسه شده است.

۵-۱- مجموعه‌داده

مجموعه‌دادهٔ مورد استفاده در این مقاله، مجموعه‌دادهٔ 20 newsgroups [1] است. این مجموعه شامل بیست گروه از پیش تعیین شده است که هرکدام یک گروه خبری محسوب شده و به همین دلیل می‌توان هر یک از آنها را یک خوشه از اسناد دانست که توسط کاربر انسانی تشکیل شده است. در مقاله حاضر، از این دسته‌ها برای ارزیابی عملکرد مدل پیشنهادی در خوشه‌بندی اسناد استفاده شده است. از هریک از این دسته‌ها دویست سند، به‌طور تصادفی انتخاب شده است که صدتای آنها به آزمون و صدتای دیگر به آموزش مدل اختصاص یافته‌اند. به این ترتیب دوهزار سند برای آموزش و دوهزار سند برای آزمون استفاده شده است. ایست‌واژه‌ها، علائم و اعداد حذف شده‌اند. نشانی‌های وب و ایمیل با @ جایگزین شده‌اند. هم‌چنین واژگان نادر که در کمتر از پنج سند در کل مجموعه تکرار شده نیز حذف شده‌اند. پس از

[1]-http://qwone.com/~jason/20Newsgroups

[DOI: 10.29252/jsdp.15.4.57]

پیش‌پردازش، این مجموعه‌داده دارای ۴۲۵۰۳۶ واژه است که بر مبنای مجموعه‌واژه‌ای ۱۸۹۳۱تایی تشکیل شده‌اند به این ترتیب، میانگین طول اسناد حدود ۱۰۶ واژه است.

۲-۵- آزمایش‌ها

برای تمامی آزمایش‌های زیر، تنظیمات یادشده در جدول (۱) را استفاده کرده‌ایم. تعداد موضوعات در این جدول ذکر نشده‌اند؛ زیرا نتایج آزمایش‌ها برای تعداد موضوعات مختلف گزارش شده‌اند. اندازه فراپارامترها مشابه مقادیری است که به‌صورت قراردادی در بسیاری از مدل‌های موجود مورد استفاده قرار گرفته است. این پارامترها نوعی اثر هموارسازی بر روی توزیع‌های چندجمله‌ای دارند. برای اطلاعات بیشتر در زمینه اثر این پارامترها در مدل‌های احتمالاتی دریکله-چندجمله‌ای می‌توان به [22] مراجعه کرد.

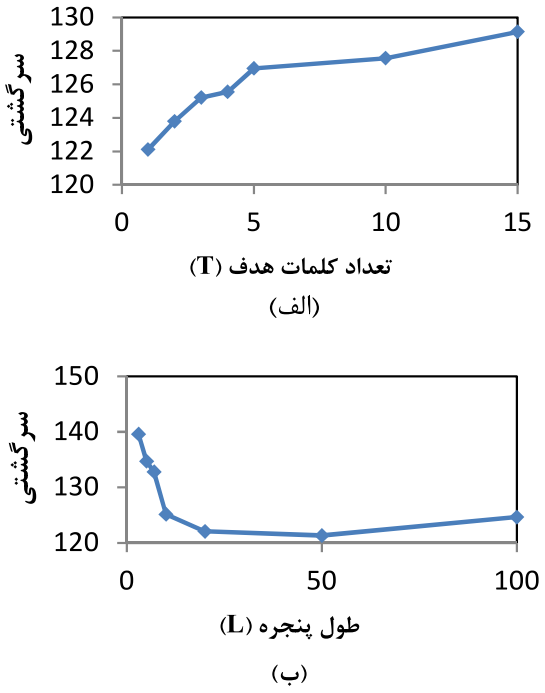
(جدول-۱):تنظیمات مورد استفاده در آزمایش‌ها

(Table-1): experimental settings

پارامتر	مقدار
β	0.01
α_{LLDA}	$50/(K + 20 * L)$
$\alpha_{LDA,BTM}$	$50/K$
Number of iterations	1000
Burn-in	200
lag	100
L	20
P	10

فراپارامترها در مدل پیشنهادی نیز به‌طور مشابه تعریف شده‌اند. برای مدل پیشنهادی به‌دلیل این‌که علاوه‌بر تعداد موضوعات K، طول پنجره L نیز یک پارامتر مؤثر است، مقدار آن در تعیین اندازه فراپارامترها مورد توجه قرار گرفته است. چندبرابرکردن L به این منظور است که در مقابل K بی‌اثر نباشد.

چنانکه ذکر شد، برای محاسبه پارامترهای مدل سعی می‌کنیم مقادیر آنها را به‌گونه‌ای تعیین کنیم که احتمال بیشتری به واژگان مجموعه اختصاص یابد. بر این اساس به‌طورمعمول یکی از روش‌هایی که برای ارزیابی مدل‌های موضوعی مورد استفاده قرار می‌گیرد معیار سرگشتگی (پرپلکسیتی) است. با این حال، در سال‌های اخیر ثابت شده است که سرگشتگی، معیار چندان قابل اعتمادی برای مقایسه و سنجش مدل‌های موضوعی نیست [23]. در این مقاله، از سرگشتگی تنها برای انتخاب متغیرهای مدل استفاده شده است. این معیار به‌صورت معکوس میانگین هندسی احتمال واژگان مجموعه، محاسبه می‌شود. احتمال یادشده نیز، بر اساس مدل موضوعی به‌دست می‌آید. نکته دیگری که در اینجا لازم به ذکر است، میزان هم‌پوشانی پنجره‌ها یا به‌عبارتی تعداد واژگان هدف در هر پنجره است که با T نشان داده می‌شود. (شکل-الف) مقدار سرگشتگی را بر روی مجموعه‌دادهٔ یادشده و تعداد واژگان هدف از یک تا پانزده را نشان می‌دهد. در این آزمایش از تنظیمات یادشده در جدول (۱) استفاده شده است. در این شکل می‌بینیم که کم‌ترین سرگشتگی را برای مقدار $T=1$ به‌دست آورده‌ایم. بنابراین تمامی مطالب از این پس با این فرض بیان می‌شوند که T برابر یک است. شکل (۶-ب) مقدار سرگشتگی را بر روی مجموعه‌دادهٔ یادشده و برای اندازهٔ پنجره‌های مختلف نشان می‌دهد. بقیهٔ تنظیمات در این آزمایش مطابق تنظیمات یادشده در جدول (۱) است. همان‌طور که می‌بینید کمترین مقدار سرگشتگی را در اندازهٔ پنجره بیست داریم. تا اندازهٔ بیست، با افزایش اندازهٔ پنجره، مدل پیشنهادی داده‌ها را بهتر و بهتر مدل می‌کند ولی اگر اندازهٔ پنجره بیشتر از بیست شود، این روند کند و سپس معکوس خواهد شد.



(شکل-۶): تغییرات سرگشتگی به‌عنوان تابعی از تعداد عناصر هدف در هر پنجره (الف) و طول پنجره‌های هم‌پوشان (ب)

(Figure-6): Perplexity as a function of the number of target words in each window (right) and window length (left)

در جدول (۲) تعدادی از موضوعات تولیدشده با مدل پیشنهادی را مشاهده می‌کنید. این موضوعات از بین دویست موضوع تولیدشده توسط مدل پیشنهادی انتخاب شده‌اند. موضوع ۴۷ در جدول (۲)، پرتکرارترین موضوع در نمونه سندی از مجموعه‌داده است که در شکل (۷) نمایش داده شده است. در این شکل واژگانی که به این موضوع اختصاص یافته‌اند با زمینهٔ رنگی مشخص شده‌اند. برچسب‌هایی که به

[DOI: 10.29252/jsdp.15.4.57]

واژگان داده شده ،است موضوع منتسب به آنها را نمایش می‌دهد. واژگانی که برچسب ندارند، واژگانی هستند که در جریان پیش‌پردازش، حذف شده‌اند. همان‌طور که می‌بینید، دنباله‌های پشت‌سرهم واژگان به یک موضوع اختصاص یافته‌اند.

در جدول (۲) و جدول (۳)، نمونه‌هایی از موضوعات تولیدشده با LDA و BTM نمایش داده شده‌اند. این نمونه‌ها به‌گونه‌ای انتخاب شده‌اند که موضوعات مربوط به آنها بین هر سه روش، مشترک باشند. مدل‌های LDA و BTM کمترین سرگشتگی را برای صد موضوع نشان داده‌اند. بنابراین پنج موضوع زیر از بین صد موضوع تولیدشده توسط آنها با تنظیمات یادشده در جدول (۱) انتخاب شده‌اند. موضوع ۱۹۲ از مدل پیشنهادی را در نظر بگیرید. این موضوع که در مورد یهودیان است، متناظر با موضوع ۵۹ از LDA است. موضوع ۵۹ حاوی واژگانی مانند "common" و "year" و "important" است که نه‌تنها واژگانی عمومی بوده و می‌توانند نمایندهٔ موضوعات بسیاری باشند، بلکه درفهرست مرتب واژگان قبل از واژگان خاص‌تر و مرتبط‌تری مانند " religious" و "christian" قرار گرفته‌اند.

(جدول-۲): نمونه هایی از موضوعات تولیدشده با مدل پیشنهادی برای مجموعه‌داده مورد استفاده، تحت تنظیمات موجود در جدول (۱)

(Table-2): Samples of topics generated by the proposed model, based on the settings of table 1

47	139	145	128	178	181	192
car	god	god	water	software	israel	jews
speed	truth	spirit	air	windows	lebanon	jewish
cars	faith	jesus	plants	system	israeli	religious
engine	belief	holy son	river	unix	lebanese	state
driving	accept	father	pressure	dos	civilians	muslim
road	evidence	christ	exhaust	pc	villages	conflict
front	exist	man	heavy	run	soldiers	society
miles	gods	nature	affect	support	peace	christian
drive	reliable	eternal	cooling	hardware	attacks	ottoman
left	means		temperature	network	syria	israel

-Zero to very fast[47] very quickly[47]... lastest rumor[153] is 115 hp[47] at the rear[47] wheel[47], handles[47] like a dream[104] in a straight[189] line[138] to 80-100, and then gets a tad[82] upset[153] according to a review[82] in Cycle[138] World[189]... cornering[189], er[104] well, you can't have everything...
-Sure you can have everything, if by "everything" you mean fast[47] straight[47] line[133] performance[47] AND handling[47] – present[190] day[47] liter[47] sport[196] bikes[47] have more horsepower[47] and have faster[47] 0-60 and 1/4 mile[47] times[190] than the V-max... Plus, they corner[47] just a bit[190] better...
-Seriously, handling[47] is probably as good[190] as the big[196] standards[190] of the early[196] 80's but not compareable to whats[47] state[104] of the art[175] these days[60].

(شکل-۷): نمونه سندی که واژگان آن با موضوعات منتسب به آنها توسط مدل پیشنهادی برچسب خورده‌اند.
واژگان بدون برچسب واژگانی هستند که در جریان پیش‌پردازش حذف شده‌اند.

(Figure-7): An example document with words that are tagged by their assigned topics.
The tag-less words have been removed during the preprocessing.

(جدول-۳): نمونه‌هایی از موضوعات تولیدشده با روش LDA برای مجموعه‌داده مورد استفاده و تحت تنظیمات یاد شده در جدول (۱)

(Table-3): Samples of topics generated by LDA, based on the settings mentioned in table 1

49	34	91	37	59
car	god	water	windows	jews
cars	jesus	high	mode	jewish
engine	christian	heavy	card	common
front	christ	hot	version	university
door	christians	power	driver	religious
driving	bible	solar	work	professor
wheel	man	battery	problem	years
drive	christianity	plants	dos	important
miles	faith	john	drivers	christian
nice	spirit	air		history

[DOI: 10.29252/jsdp.15.4.57]

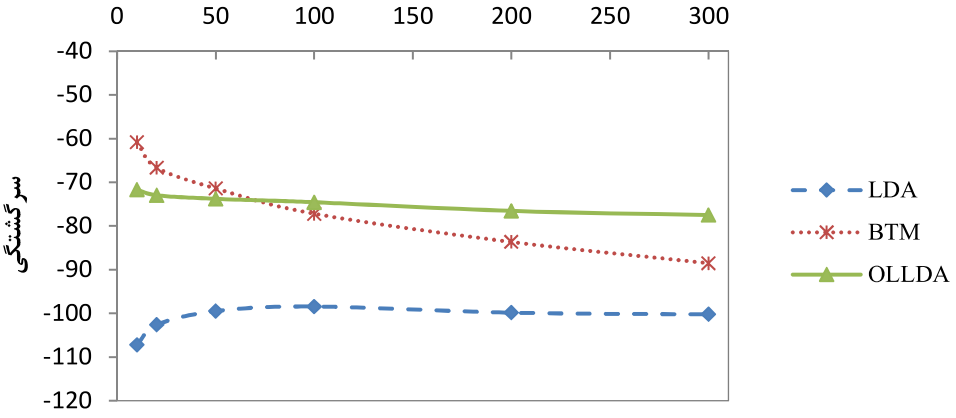
(جدول-۴): نمونه‌هایی از موضوعات تولیدشده با روش BTM برای مجموعه‌داده مورد استفاده و تحت تنظیمات یادشده در جدول (۱)

(Table-4): Samples of topics generated by BTM, based on the settings mentioned in table 1

6	39	31	67	65
@ continental people car cars fake kit good sport rear	god spirit @ holy father son people essence jesus time	@ water unit people usa aftermarket plants installed worth les	@ bus drive mb john card windows local work computer	israel lebanon israeli @ lebanese people time villages peace civilians

موضوع ۶۵ از BTM نیز که متناظر با موضوعات یادشده است، در مقایسه با موضوع ۱۴۹دارای مشکل مشابه است. واژه "@" را در نظر بگیرید؛ این واژه که جایگزین نشانی‌های وب و رایانامه در مجموعه‌داده شده است، یکی از پرتکرارترین واژگان مجموعه است و نمی‌توان آن را نمایندهٔ خاصی دانست؛ ولی در موضوع ۶۵ در چهارمین جایگاه و قبل از واژگان مهم‌تر ظاهر شده است.

حال به موضوع ۳۴ از LDA توجه کنید. این موضوع در مورد ایمان و مسیحیت است. متناظر با این موضوع برای مدل پیشنهادی موضوعات ۱۳۹ و ۱۴۵ را داریم. می‌توان گفت که موضوع ۳۴ در مدل پیشنهادی به موضوعات خاص‌تر ۱۳۹ و ۱۴۵ شکسته شده است که اولی در مورد ایمان و دومی در مورد مسیحیت است. شرایط مشابه را می‌توان در سایر موضوعات نمونه نیز مشاهده کرد.



(شکل -۸): مقدار انسجام موضوعات تولید شده توسط مدل پیشنهادی ، LDA و BTM برای تعداد موضوعات مختلف و بر اساس تنظیمات یادشده در جدول (۱)

(Figure-8): Topic coherences of BTM, LDA and the proposed model, based on the settings mentioned in table 1

بسیاری از موضوعاتی که مدل‌های موضوعی تولید می‌کنند از دید ناظر انسانی بی‌معنا یا بسیار کلی هستند. تعداد این نوع موضوعات با افزایش تعداد موضوعات تولیدشده بیشتر و بیشتر می‌شود [24]. انسجام، معیاری است که برای سنجش کیفیت و معناداری موضوعات تولیدشده با مدل‌های موضوعی معرفی شده است. در این مقاله از یکی از پرکاربردترین روش‌هایی که در سال‌های اخیر برای محاسبهٔ انسجام موضوعات پیشنهاد شده است [25] استفاده کرده‌ایم. نتایج این روش بسیار منطبق با قضاوت انسان هستند. روشن است که هرچه موضوعی منسجم‌تر باشد موضوع بهتری محسوب می‌شود.

۱-۲-۵- انسجام موضوعات

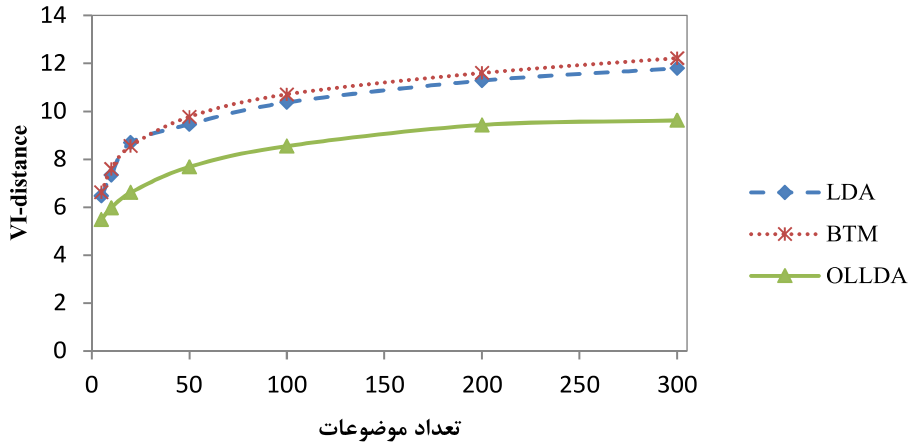
انسجام موضوعات تولیدشده با مدل‌های مورد توجه این مقاله را برای تعداد موضوعات مختلف نشان می‌دهد. این شکل آنچه را در نمونه موضوعات یادشده در بخش قبل به‌صورت کیفی

[DOI: 10.29252/jsdp.15.4.57]

دیدیم با ارزیابی کمی تأیید می‌کند. در این شکل می‌بینید که روش پیشنهادی دارای بیشترین انسجام است. از طرفی، هرچند همان‌طور که توقع داشتیم، انسجام با افزایش تعداد موضوعات در هر سه روش کاهش یافته است، این کاهش در روش پیشنهادی خفیف‌تر است. انسجام موضوعات تولیدشده با BTM در ابتدا بالاتر از موضوعات تولیدشده با مدل پیشنهادی است. ولی با افزایش تعداد موضوعات، انسجام آنها در BTM به‌شدّت کاهش می‌یابد.

۲-۲-۵- خوشه‌بندی اسناد

همان‌طور که پیش از این گفته شد، توقع داریم مدل پیشنهادی به دلیل استفاده از روابط محلی واژگان، قادر باشد موضوعاتی را استخراج کند که معنادارتر و منسجم‌تر هستند. بنابراین می‌توانیم انتظار داشته باشیم خوشه‌بندی‌ای که با مدل پیشنهادی تولید می‌شود نیز معنی‌دارتر و به قضاوت انسان شبیه‌تر باشد. برای بررسی این مسئله خوشه‌بندی تولیدشده با سه مدل مورد توجه، این مقاله را با گروه‌های خبری مربوط به اسناد مجموعه‌داده مقایسه کرده‌ایم. در مجموعه‌دادهٔ مورد استفاده، هر سند فقط متعلّق به یکی از این گروه‌های خبری است؛ بنابراین گروه‌های یادشده یک خوشه‌بندی سخت محسوب می‌شوند. برای مقایسه از معیاری به نام Variation of Information distance (VI-distance) [26] استفاده کرده‌ایم. روش VI-distance قادر است خوشه‌بندی‌هایی را با تعداد خوشه‌های متفاوت، با هم مقایسه کند؛ یعنی محدود به یکسان‌بودن تعداد خوشه‌ها نیست. هم‌چنین محدود به هم‌نوع‌بودن خوشه‌ها هم نیست. یعنی قادر است، یک خوشه‌بندی نرم را با یک خوشه‌بندی سخت مقایسه کند. همان‌طور که پیش از این گفته شد، در مدل‌های LDA و BTM، هر سند به‌صورت یک توزیع بر روی موضوعات نمایش داده می‌شود. از این توزیع‌ها می‌توان به‌عنوان یک خوشه‌بندی نرم برای اسناد استفاده کرد. در مدل پیشنهادی چنین توزیعی بر روی هر پنجره تعریف می‌شود. بنابراین برای به‌دست‌آوردن خوشه‌بندی اسناد می‌توان در هر سند، میانگین مقادیر به‌دست‌آمده را برای هر موضوع در پنجره‌های مختلف محاسبه کرد.



(شکل-۹): مقدار فاصله کلاسترهای تولیدشده با مدل پیشنهادی ، LDA و BTM از دسته‌های ساخته‌شده توسط انسان، برای تعداد موضوعات مختلف و بر اساس تنظیمات یادشدهٔ جدول (۱)

(Figure-9): Clustering distances of BTM, LDA and the proposed model with the human generated categories, for several numbers of topics and based on the settings mentioned in table 1

در این مقاله، از خوشه‌بندی‌های یادشده برای سنجش توانایی مدل‌ها در تولید خوشه‌بندی منطبق بر قضاوت انسان، استفاده شده است. حاصل این مقایسه در شکل (۹) برای تعداد موضوعات (خوشه‌های) مختلف، نمایش داده شده است. همان‌طور که در شکل قابل مشاهده است، خوشه‌های تولیدشده با روش پیشنهادی در تمام موارد کم‌ترین فاصله را با گروه‌های خبری دارند.

۶- جمع‌بندی

در این مقاله، یک مدل موضوعی احتمالاتی جدید معرفی شده است که در آن، هر سند، مجموعه‌ای از پنجره‌های هم‌پوشان محسوب می‌شود. هر پنجره متناظر با یکی از واژگان سند است. واژه یادشده، واژهٔ هدفِ آن پنجره نامیده می‌شود. هر پنجره یک توزیع بر روی موضوعات است و موضوع واژه هدف

[DOI: 10.29252/jsdp.15.4.57]

از این توزیع استخراج می‌شود. به‌دلیل هم‌پوشانی پنجره‌ها، موضوع هر واژه بر واژگان اطرافش اثر می‌گذارد. هرچه دو واژه دورتر از یکدیگر باشند در پنجره‌های کمتری هم‌رخدادند و بنابراین اثر کمتری بر یکدیگر دارند. به این ترتیب، مدل پیشنهادی به‌صورت ضمنی از اطلاعات ترتیب واژگان که در فاصله آنها منعکس می‌شود، بهره می‌برد؛ ولی وابسته به ترتیب دقیق واژگان نیست و وابستگی صریحی بین واژگان در آن تعریف نشده است. به همین دلیل برخلاف مدل‌های موضوعی مبتنی بر n-gram مانند BTM، تُنُکی برای مدل پیشنهادی یک چالش اساسی محسوب نمی‌شود. از طرفی، برخلاف مدل‌هایی مانند LDA که مبتنی بر روابط واژگان در سطح یک سند هستند، قادر است از ارتباطات محلی واژگان که بر اساس اصل هم‌رخدادی نماینده بهتری برای ارتباطات معنی‌دار واژگان هستند، بهره بگیرد. این مدل بر روی یک زیرمجموعهٔ چهارهزار سندی از مجموعه‌داده 20 newsgroup با مدل‌های یادشده مقایسه شده است. بر اساس نتایج این آزمایش‌ها، مدل پیشنهادی موضوعات منسجم‌تری تولید می‌کند. این نتیجه در چند مثال نیز نمایش داده شده است. مدل پیشنهادی در کاربرد خوشه‌بندی اسناد نیز با دو روش LDA و BTM مقایسه شده است. بر اساس نتایج این مقایسه، مدل قادر است، خوشه‌هایی بسازد که بیشتر از خوشه‌های ساخته‌شده با دو روش یادشده، به خوشه‌های ساخته‌شده توسط انسان شبیه هستند.

۷- مراجع 7- References

[۱]فیلی هشام، قادر حمیدرضا، آنالویی مرتضی. یک مدل بیزی برای استخراج باناظر گرامر زبان طبیعی. پردازش علائم و داده‌ها. ۱۳۹۱; ۹ (۱) :۳۴-۱۹

[1] Faili, H., H. Ghader, and M. Morteza Analoui, "A Bayesian Model for Supervised Grammar Induction," *Signal and Data Processing*, 2012. 9(1), pp. 19-34.

[2] D., et al. Wang, "Multi-document summarization using sentence-based topic models," 2009. *Association for Computational Linguistics*.

[۳] صادقی سیده ساره، وزیرنژاد بهرام. خلاصه‌ساز متون روایی مبتنی بر جنبه‌های شناختی ذهن انسان. پردازش علائم و داده‌ها. ۱۳۹۴; ۱۲ (۲) :۹۶-۸۷

[3] S. S. Sadegi and B, vazir nejad, "Extractive summarization based on cognitive aspects of human mind for narrative text," *Signal and Data Processing*, vol.12(2), pp. 87-96, 2015

[4] H. Zhang and G. Zhong, "Improving short text classification by learning vector representations of both words and hidden topics," *Knowledge-Based Systems*, 2016. 102: pp. 76-86.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, pp. 993-1022, 2003.

[6] H.M. Wallch, "Topic modeling: beyond bag-of-words," *ACM*, 2006.

[7] C.D. Manning, et al., "Introduction to Information Retrieval," *Cambridge University Press*, pp. 496, 2008.

[8] im Walde, S.S. and A. Melinger, "An in-depth look into the co-occurrence distribution of semantic associates," *Italian Journal of Linguistics, Special Issue on From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science*, 2008.

[9] N. Barbieri, et al., "Probabilistic topic models for sequence data," *Machine learning*, vol.93(1), pp. 5-29, 2013.

[10] T.L. Griffiths, M. Steyvers, and J.B. Tenenbaum, "*Topics in semantic representation*." *Psychological review*, vol.114(2), pp. 211, 2007.

[11] X. Wang, A. McCallum, and X. Wei. "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," *IEEE*, 2007.

[12] G. Yang, et al., "A novel contextual topic model for multi-document summarization, "*Expert Systems with Applications*, vol. 42(3), pp. 1340-1352, 2015.

[13] S. Jameel, W. Lam, and L. Bing, "Supervised topic models with word order structure for document classification and retrieval learning," *Information Retrieval Journal*, vol.18(4), pp. 283-330, 2015.

[14] Y.W. The, "A hierarchical Bayesian language model based on Pitman-Yor processes," *Association for Computational Linguistics*, 2006.

[15] H. Noji, D. Mochihashi, and Y. Miyao. "Improvements to the Bayesian Topic N-Gram Models," in *EMNLP*, 2013.

[16] I. Sato and H. Nakagawa. "Topic models with power-law using Pitman-Yor process," *ACM*, 2010.

[17] Y.-S. Jeong and H.-J. Choi, "Overlapped latent Dirichlet allocation for efficient image segmentation," *Soft Computing*, vol. 19(4), pp. 829-838.

[18] Y. Zue, J. Zhao, and K. Xu, "Word network topic model: a simple but general solution for short and

[DOI: 10.29252/jsdp.15.4.57]

imbalanced texts," *Knowledge and Information Systems*, pp. 1-20, 2014.

[19] W. Ou, Z. Xie, and Z. Lv. "Spatially Regularized Latent topic Model for Simultaneous object discovery and segmentation," in *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*. 2015. IEEE.

[20] T.L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National academy of Sciences*, 2004. 101(suppl 1), pp. 5228-5235.

[21] T. Minka and J. Lafferty. "Expectation-propagation for the generative aspect model," *Morgan Kaufmann Publishers In*, 2002.

[22] J. Rennie, 20 Newsgroups. Available from: http://qwone.com/~jason/20Newsgroups/20news-18828.tar.gz

[23] G. Heinrich, "Parameter estimation for text analysis," University of Leipzig, Tech. Rep, 2008.

[24] D. Newman, et al. "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2010. Association for Computational Linguistics.

[25] D. O'Callaghan, et al., "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42(13), pp. 5645-5657, 2013.

[26] D. Mimno , et al. "Optimizing semantic coherence in topic models," *Association for Computational Linguistics*, 2011.

[27] M. Meilă, "Comparing clusterings by the variation of information, in Learning theory and kernel machines," Springer, 2003, pp. 173-187.

مرضیه رحیمی فارغ‌التحصیل مقطع دکترای تخصصی رشته مهندسی کامپیوتر، گرایش هوش مصنوعی از دانشگاه صنعتی شاهرود است. وی مقاطع کارشناسی و کارشناسی ارشد را نیز در همان دانشگاه به پایان رسانیده و در حال حاضر استادیار دانشگاه صنعتی شاهرود است. زمینه‌های پژوهشی مورد علاقهٔ ایشان یادگیری آماری، داده‌کاوی و مدل‌سازی موضوعی است.

نشانی رایانامه ایشان عبارت است از:

marziea.rahimi@sharoodut.ac.ir

مرتضی زاهدی دارای دکترای تخصصی کامپیوتر از دانشگاه RWTH-Aachen آلمان است. ایشان مدارک کارشناسی ارشد و کارشناسی خود را به‌ترتیب از دانشگاه‌های تهران و صنعتی امیرکبیر اخذ کرده و در حال حاضر عضو هیأت علمی دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی شاهرود است. زمینه‌های پژوهشی مورد علاقهٔ ایشان، تعامل انسان و کامپیوتر، شناسایی الگو، پردازش تصویر و ویدئو، و بینایی ماشین، با تأکید بر روش‌های مبتنی بر اطلاعات و دانش آماری است.

نشانی رایانامه ایشان عبارت است از:

zahedi@sharoodut.ac.ir

هدی مشایخی فارغ‌التحصیل مقطع دکترای تخصصی رشته مهندسی کامپیوتر، گرایش نرم‌افزار از دانشگاه صنعتی شریف است. پیش از آن، مقاطع کارشناسی و کارشناسی ارشد را نیز در همان دانشگاه به پایان رسانیده است. وی در حال حاضر استادیار دانشگاه صنعتی شاهرود است. داده‌کاوی و یادگیری در زمینه داده‌های حجیم، و پردازش توزیع‌شده از جمله علایق پژوهشی ایشان است.

نشانی رایانامه ایشان عبارت است از:

hmashayekhi@sharoodut.ac.ir

[DOI: 10.29252/jsdp.15.4.57]