

**مدل نظارت شده‌ی ضعیف برای تشخیص مشترک احساس-موضوع از داده‌های متنی**

**مسعود فاطمی**

گرایش هوش مصنوعی و رباتیک/ رشته‌ی مهندسی کامپیوتر

دانشکده‌ی برق و کامپیوتر

9314994

گزارش پروژه‎‌ی پایانی درس یادگیری ماشین

دکتر پالهنگ

زمستان 94

**چکیده**

**هدف از تحلیل احساس یا کاوش عقاید استفاده از ابزارهای خودکار برای تشخیص و استخراج اطلاعات مفهمومی مانند عقاید، نگرش‌ها، و احساسات بیان شده در متن می‌‌باشد. در این پروژه یک ساختار احتمالاتی جدید تحت عنوان مدل مشترک احساس‌-‌‌موضوع بر اساس مدل تخصیص دیریکله‌ی پنهان معرفی‌ و پیاده سازی می‌‌شود، که به صورت همزمان موضوع و احساس را از داده‌های متنی استخراج می‌‌کند. در کنار مدل‌های نظارت شده که برای تشخیص احساس و طبقه‌بندی متون براساس احساس وجود دارد و در بیشتر موارد استفاده از چنین مدل‌هایی در یک دامنه‌ی دیگر منجر به نتایج نامطلوب می‌گردد مدل معرفی‌ شده در اینجا یک مدل نظارت شده‎‌ی ضعیف می‌‌باشد که این خاصیت باعث می‌‌گردد اجرای مدل بر روی دامنه‌های دیگر منجر به تولید خروجی مطلوب گردد و کاهش کارایی در خروجی مدل به وجود نیاید. مدل نهایی پیاده‌سازی شده در اینجا بر روی ‌یک پایگاه داده مورد آزمایش قرار می‌گیرد و نتایج حاصل با توجه به معیار سرگشتی مورد ارزیابی قرار می‌‌گیرند.**

***کلمات کلیدی: مدل موضوعی، مدل تخصیص دریکله‌ی پنهان، مدل مولد، تقریب تغییرات، توزیع پسین***

1. **مقدمه**

در این پروژه هدف تشخیص احساس از داده‌های متنی می‌‌باشد. مدلی‌ که در اینجا برای هدف مورد نظر پیاده‌سازی شده است بسط داده شده‌ی ساختار تخصص دیریکله‌ی پنهان[[1]](#footnote-1) می‌‌باشد که از آن برای استخراج همزمان احساس[[2]](#footnote-2) و موضوع از داده‌های متنی استفاده می‌‌شود. مدل تخصیص دریکله‌ی پنهان که در ادامه به صورت کامل شرح داده می‌‌شود یک مدل موضوعی[[3]](#footnote-3) می‌‌باشد. با توجه به تعریفی‌ که برای مدل‌های موضوعی وجود دارد چنین ساختارهایی به عنوان زیر مجموعه‌ای از مدل‌های احتمالاتی مولد[[4]](#footnote-4) در نظر گرفته می‌‌شوند. با اجرای یک مدل موضوعی بر روی مجموعه‌ای از اسناد یک چکیده از موضوع‌های موجود در سند‌های مختلف به عنوان خروجی دریافت می‌‌شود که بسته به کاربرد مورد نظر می‌‌توان آن را به عنوان یک نمایش منسجم از مجموعه اسناد در نظر گرفت. یا در کاربردی دیگر این مجموعه موضوع‌های موجود در متن را می‌‌توان به عنوان یک بردار ویژگی‌ در نظر گرفت و برای مثال در فرآیند طبقه‌بندی[[5]](#footnote-5) متون از آن استفاده کرد.

همان‌طور که ذکر شد مدلی‌ که در اینجا بسط داده می‌‌شود گسترش یافته‌‌ی مدل تخصیص دیریکله‌ی پنهان می‌‌باشد که علاوه بر موضوع، احساس موجود در یک مجموعه سند که می‌‌تواند مجموعه‌ای از بازبینی‌ها باشد را برای ما استخراج می‌‌کند.  
همان‌طور که در پیشنهاد پروژه که پیش از این اعلام گردیده بود برای پیاده سازی این مدل از نرم افزار متلب استفاده شده و کد نوشته شده در محیط این برنامه قابل اجرا می‌‌باشد. در ادامه با خروجی‌ها، ورودی‌ها، بخش‌های مختلف پیاده سازی و نحوه‌ی اجرا به صورت کامل آشنا می‌‌شویم.

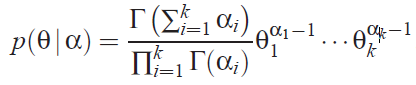
1. **تخصیص دیریکله‌ی پنهان**

همان‌طور که پیش از این نیز ذکر شد تخصیص دیریکله‌ی پنهان یک مدل احتمالاتی بر روی مجموعه‌ای از اسناد می‌‌باشد. فرضی‌ که در این مدل می‌‌شود بر پایه‌ی این است که هر سند متنی را می‌‌توان به صورت یک توزیع مخلوط[[6]](#footnote-6) بر روی موضوع‌های مختلف‌ که در آن هر موضوع به وسیله‌ی یک توزیع بر روی کلمه‌ها مشخص می‌‌شود در نظر گرفت[1].

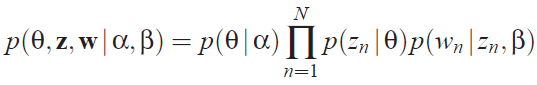
با توجه با توضیحات بالا مدل LDA فرآیند مولد زیر را برای هر سند **w** در مجموعه سندD در نظر می‌‌گیرد[1]:

1. انتخاب N از یک توزیع پواسن مانند £.  
 2. انتخاب θ از یک توزیع دیریکله با بردار α.  
 3. برای هر N کلمه مانند   
   (الف) انتخاب یک موضوع از توزیع چندجمله‌ای[[7]](#footnote-7) θ.  
    (ب) انتخاب یک کلمه مانند از که یک توزیع احتمالی شرطی و مشروط به موضوع می‌‌باشد.

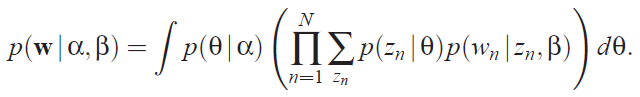
یک متغیر تصادفی دیریکله‌ی k-بعدی مانند θ با توجه به رابطه‌ی توزیع دیریکله به فرم زیر می‌‌باشد و چگالی آن از این رابطه بدست می‌آید[1]:

(1) 

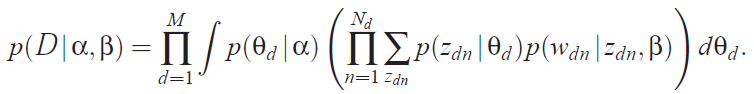
که در آن پارامت یک بردار-بعدی می‌‌باشد که در آن هر ، و تابع گاما[[8]](#footnote-8) می‌‌باشد. با توجه به پارامتر‌های محاسبه شده‌ی α و توزیع مشترک مخلوط موضوع‌ها یا ، یک مجموعه‌ از N موضوع **z** و یک مجموعه از N کلمه‌ی **w** از رابطه‌ی زیر محاسبه می‌‌شوند[1]:

(2)

که در آن به بیانی ساده امین موضوع از بردار احتمالاتی می‌‌باشد. با انتگرال‌گیری از رابطه‌ی 2 بر روی و محاسبه سیگما نسبت به z آنچه که حاصل می‌‌شود توزیع حاشیه‌ای[[9]](#footnote-9) برای یک سند می‌‌باشد که فرمی به شکل زیر دارد[1]:

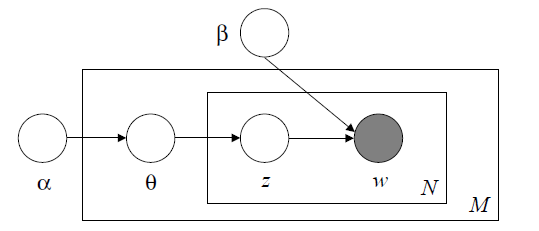
(3)

در نهایت با ضرب کردن این احتمال‌های حاشیه‌ای برای هر سند در یکدیگر احتمال مورد نظر برای یک مجموعه سند خاص را می‌‌توان بدست آورد[1]:

(4)

لازم به ذکر می‌‌باشد که ماتریس که در روابط بالا موجود می‌‌باشد ماتریسی با ابعاد می‌‌باشد که در آن تعداد کلمات متمایز موجود در مجموعه سند یا به عبارت دیگر تعداد کلمه‌های موجود در لغات‌نامه و تعداد موضوع‌های نهان[[10]](#footnote-10) خواسته شده می‌‌باشد. با توجه به این تعریف عناصر ماتریس برابر احتمال وجود هر کلمه در هر موضوع می‌‌باشد. برای مثل نشان دهنده‌ی احتمال مشاهده‌ی کلمه‌ی iام در موضوع jام می‌‌باشد.

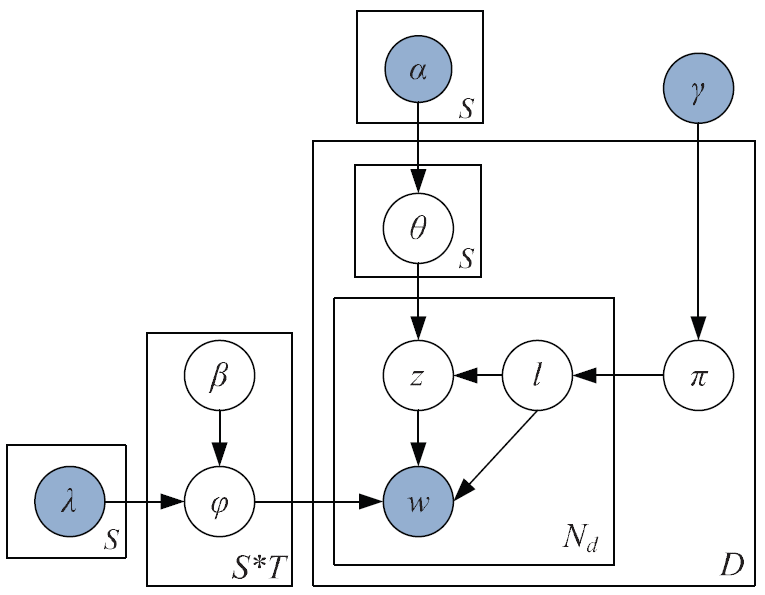
مدل LDA که روابط آن در اینجا شرح داده شد به عنوان یک مدل گرافی احتمالی[[11]](#footnote-11)‌ در شکل ۱ نشان داده شده است. همان طور که در شکل ۱ نیز مشخص می‌‌باشد LDAیک مدل ۳ سطحی می‌‌باشد. پارامتر‌های α و پارامتر‌های در سطح مجموعه سند می‌باشند و در کل فرآیند تولید یک مجموعه سند تنها یک بار نمونه‌برداری می‌‌شوند و تا انتهای کار ثابت هستند. متغیر یک متغیر در سطح سند می‌باشد، یا به بیان دیگر برای هر سند یک بار نمونه‌برداری می‌‌گردد. و در انتها متغیر‌های و متغیر‌های در سطح کلمه می‌‌باشند که که برای هر کلمه در هر سند یک بار نمونه‌برداری می‌‌شوند. لازم به ذکر می‌‌باشد که در شکل‌هایی مانند آنچه که در اینجا به عنوان مدل LDA نشان داده شده است و به آن‌ها مدل گرافی گفت می‌شود، پیکان‌ها نشان دهنده‌ی وابستگی، مستطیل‌ها نشان دهنده‌ی تکرار و دایره‌ها نشان دهنده‌ی متغیر‌های تصادفی می‌‌باشند. البته خود این متغیرهای تصادفی بر دو نوع پنهان (سفید) یا مشاهده‌شده[[12]](#footnote-12) (خاکستری) می‌‌باشند.

شکل 1

1. **مدل مشترک احساس-موضوع**

با توجه به آنچه که در بخش‌های قبل گفت شده مدل مشترک احساس-موضوع[[13]](#footnote-13) که در شکل 2 مشاهده می‌‌شود همانند یک مدل احتمالاتی مولد می‌‌باشد. در واقع گسترش یافته‌ی‌ مدل می‌‌باشد که علاوه بر تشخیص موضوع به تشخیص احساس از داده‌های متنی نیز می‌پردازد. مدل که در قسمت قبل شرح داده شد بر این فرض می‌‌باشد که هر سند به صورت مخلوطی از موضوع‌ها می‌‌باشد، در حالی‌ که هر موضوع یک توزیع احتمالاتی بر روی کلمه‌ها می‌‌باشد. برای مدل کردن احساس در JST یک لایه بین سند و موضوع اضافه می‌کنیم. با این وجود یک مدل چهار لایه می‌‌باشد که در آن اسناد با احساسات همراه هستند و احساسات با موضوع‌ها همراه هستند و در نهایت کلمات با هردو برچسب احساس و موضوع همراه هستند.

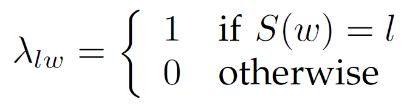
فرض کنید که یک مجموعه سند در اختیار داریم که شامل سند می‌‌باشد که به صورت مشخص می‌‌شوند، هر سند موجود دنباله‌ای از کلمه می‌‌باشد که با مشخص می‌‌شوند و هر کلمه در داخل سند یک مورد از کلمه‌ی مجزای واژگان می‌باشد که به صورت شاخص‌گذاری شده‌اند. همچنین تعداد برچسب‌های احساس متمایز می‌‌باشد و برابر است با تعداد کلی‌ موضوع‌ها. روند تولید کلمه‌ی در سند تحت شامل سه‌ مرحله می‌‌شود. ابتدا یک برچسب احساس مثلا از که توزیع احتمالاتی احساس برای هر سند می‌‌باشد انتخاب می‌کنیم. سپس با توجه به برچسب احساس انتخاب شده و مشروط به آن یک موضوع از توزیع موضوعی انتخاب می‌کنیم. در نهایت یک کلمه از توزیع مشروط به هر دو برچسب موضوعی و احساس برای هر مجموع سند می‌‌کشیم و یا تولید می‌کنیم. لازم به ذکر می‌‌باشد که در مدل ما به ازای هر سند یک توزیع احتمالاتی برای موضوع‌ها داشتیم، اما در مدل به ازای هر برچسب احساس ما یک توزیع احتمالاتی برای موضوع‌ها در سند خواهیم داشت. به طور مثل اگر ما سه‌ برچسب احساس مثبت،منفی‌ یا بی‌طرف داشته باشیم، برای هر سند ما سه‌ توزیع احتمالاتی موضوعی خواهیم داشت که هر کدام متناظر با یک احساس خواهند بود. تفاوت دیگری که بین LDA و JST می‌‌باشد از این قرار است که در JST به هنگام نمونه‌برداری از کلمه، این کار تنها مشروط به توزیع موضوعی برای مجموعه سند انجام می‌‌شود ولی‌ در JST این نمونه برداری مشروط به هر دو توزیع موضوعی و احساسی‌ انجام می‌‌شود[2].

شکل2

تعریف رسمی از فرآیند مولد تحت مدل که با شکل ۲ متناظرر است به شکل زیر می‌‌باشد[2]:

1. برای هر برچسب احساس   
   - برای هر موضوع یک از توزیع می‌کشیم.
2. برای هر سند یک توزیع مانند از بدست می‌آوریم.
3. برای هر برچسب احساس در سند یک توزیع مانند از Dir(انتخاب می‌کنیم.
4. برای هر کلمه‌ی در سند   
       - یک برچسب احساس مانند از توزیع چند جمله‌ای انتخاب می‌‌کنیم،  
       - یک موضوع مانند از توزیع چند جمله‌ای انتخاب می‌‌کنیم،  
       - یک کلمه مانند از که یک توزیع چندجمله‌ای روی کلمه‌ها و مشروط به موضوع و برچسب احساس می‌‌باشد انتخاب می‌‌کنیم.
5. **ترکیب کردن دانش اولیه در مدل**

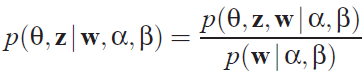
همان‌طور که پیش از این نیز ذکر شد مدل گسترش یافته‌‌ی مدل می‌‌باشد. در مقایسه با در کنار اضافه شدن یک لایه‌ی مولد از احساس همچنین یک وابستگی دیگر از با توجه به ماتریس که اندازه‌ی آن می‌‌باشد به مدل اضافه کرده‌ایم که برای ما اطلاعات اولیه‌‌ی[[14]](#footnote-14) ما در مورد احساس کلمه‌ها را به صورت کدگذاری شده به سیستم وارد می‌‌کند. ماتریس را می‌توان به عنوان یک ماتریس انتقال[[15]](#footnote-15) در نظر گرفت که توزیع دیریکله با پارامتر را به گونه‌ای تغییر می‌‌دهد که می‌توان از آن قطبش[[16]](#footnote-16) یا جهت احساس اولیه‌ی‌ کلمات را بدست آورد. برای این منظور در ابتدا با مقدار ۱ برای تمام عناصر مقدار دهی‌ می‌‌شود. سپس برای هر کلمه‌ی در لغات‌نامه و برای هر برچسب احساس اگر در لغات نامه‌ی احساس وجود داشت مقدار به صورت زیر مقدار دهی‌ می‌‌شود[2]:

(5)

که در آن برچسب احساس اولیه‌ی کلمه‌ی در لغات نامه‌ی احساسات را بر می‌گرداند. برای مثل کلمه‌ی "عالی" با شاخص i در لغت‌نامه دارای قطبش مثبت از نظر احساسی‌ می‌‌باشد. در نتیجه بردار سطری متناظر با آن در به شکل خواهد بود. در این حالت برای هر موضوع ضرب با منجر به باقی‌ ماندن تنها می‌‌شود و و برابر با صفر می‌‌شوند. بنابرین کلمه‌ی "عالی‌" تنها از توزیع موضوع کلمه‌های مثبت از توزیع دیریکله‌ای با پا پارامتر   می‌‌تواند تولید شود.

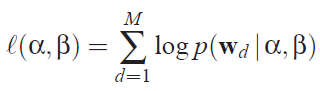
1. **استنتاج و تخمین پارامتر**

مساله‌‌ی کلیدی که در بحث استنتاج[[17]](#footnote-17) و به منظور استفاده از مدل نیاز به حل آن داریم محاسبه‌ی توزیع پسین[[18]](#footnote-18) متغیر‌های پنهان با توجه به یک سند داده شده می‌‌باشد[1]:

(6)

متاسفانه این توزیع در حالت کلی‌ قابل حل کردن نمی‌‌باشد. با توجه به اینکه محاسبه‌ی این توزیع پسین دارای استنتاج مستقیم و دقیق نیست از الگوریتم‌های استنتاجی تغیراتی[[19]](#footnote-19)‌ برای استنتاج پارامتر‌های مدل استفاده می‌‌کنند که شامل تقریب لاپلاس[[20]](#footnote-20)، تقریب تغییرات[[21]](#footnote-21)‌ و زنجیره‌ مارکوف مونت کارلو[[22]](#footnote-22) می‌‌شوند. روشی‌ که در اینجا استفاده می‌‌شود بر پایه‌ی تقریب تغییرات می‌‌باشد.  
ایده‌ی اصلی‌ در استنتاج تغیرات‌ استفاده از نامعادله‌ی جنسن[[23]](#footnote-23) می‌‌باشد که در آن یک کران پایین قابل تنظیم برای درست نمایی لگاریتم[[24]](#footnote-24) پارامتر‌ها بدست می‌‌آورند. اساسا یک خانواده از کران‌های پایین را در نظر بگیرید که به وسیله‌ی مجموعه‌ای از پارامتر‌های تغییراتی‌ شاخص‌گذاری می‌شوند. این پارامتر‌های تغییراتی‌ با استفاده از یک روش بهینه‌سازی که قصد پیدا کردن دقیق‌ترین کران پایین تا حد امکان را دارند انتخاب می‌‌شوند.

در بخش تخمین پارامتر از یک روش بیز تجربی[[25]](#footnote-25)‌ برای تخمین پارامتر‌های مدل استفاده می‌‌کنیم. به طور خاص‌تر با توجه به یک مجموعه سند داده شده به صورت ، ما تمایل به پیدا کردن و به صورتی‌ که داریم که درستنمابی لگاریتمی داده‌ها را برای ما بیشینه کند[1].

(7)

همان‌طور که پیش از این نیز ذکر شد مقدار به صورت عملی‌ قابل محاسبه نمی‌‌باشد. اگرچه با استفاده از استنتاج تغییراتی‌ می‌‌توان یک کران پایین عملی‌ برای درست نمایی لگاریتمی پیدا کرد، حال می‌‌توان این کران را نسبت به α و β بیشینه کرد. بنابراین می‌توان تخمین بیز تجربی را به صورت تقریبی برای مدل با یک روش متناوب تغییرات حداکثر انتظار[[26]](#footnote-26) پیدا کنیم. این روش در واقع کران پایین را با توجه به پارامتر‌های تغییرات‌ بیشینه می‌‌کند و سپس بعد از همگرایی این قسمت برای یک مقدار ثابت از پارامتر‌های تغیبراتی‌ این کران پایین را نسبت به پارامتر‌های مدل که α و β می‌باشند بیشینه می‌‌کند. روش شرح داده شده را می‌‌توان به صورت دو مرحله‌ی زیر نوشت:

1. (E مرحله) برای هر سند مقدار بهینه‌ی پارامترهای تغییرات را پیدا می‌‌کنیم.
2. (M مرحله) بیشینه کردن کران پایین حاصل روی لگاریتم درستنمایی با توجه به پارامتر‌های مدل که α و β می‌‌باشند.
3. **ارزیابی**

برای سنجش میزان توانایی مدل‌های موضوعی در بدست آوردن توزیع کلمات و موضوعی‌ها از معیاری تحت عنوان سرگشتگی[[27]](#footnote-27) استفاده می‌‌شود. در اینجا هم از همین معیار برای ارزیابی مدل پیاده‌سازی شده استفاده می‌‌شود. محاسبه این معیار توسط کد پیاده‌سازی شده که به پیوست تقدیم می‌گردد انجام می‌‌شود و در خروجی حاصل از اجرای برنامه نمایش داده می‌‌شود. همچنین در کد پیاده‌سازی شده از این معیار به عنوان شرط همگرایی استفاده شده است و اگر در تکرار‌های متوالی مقدار تغییر این معیار از یک حد آستانه کمتر باشد به عنوان شرط خاتمه در نظر گرفته می‌‌شود و الگوریتم همگرا می‌‌گردد.

1. **نتیجه‌گیری**

در این گزارش و پروژه یک مدل نظارت شده‌ی ضعیف برای تشخیص همزمان احساس و موضوع به صورت مشترک از داده‌های متنی را معرفی‌ و پیاده‌سازی کردیم. مدل معرفی‌ شده که JST نام دارد عمل‌کرد مناسبی برای کاربرد‌های خواسته شده از خود نشان می‌‌دهد. خاصیت نظارت شده‌ی ضعیف برای این مدل باعث می‌‌شود که در مقایسه با سایر مدل‌ها، JST به راحتی‌ قابل انتقال به یک دامنه‌ی دیگر باشد بدون کاهش محسوس در کارایی که در سایر مدل ها این اتفاق رخ می‌‌دهد. از دیگر تفاوت‌های مدل JST با سایر مدل‌ها این است که تمام مدل‌های موجود به تشخیص احساس کلی‌ متن می‌پردازند، این در حالی‌ می‌‌باشد که JST به تشخیص احساس همراه با هر موضوع و احساس کلی‌ متن می‌‌پردازد که در مسائل پیش‌رو دید بهتری را به کاربر خواهد داد.

1. **پیوست**

مدل شرح داده شده در این گزارش در نرم افزار متلب پیاده‌سازی گردیده و فایل‌های آن به پیوست این گزارش تقدیم می‌‌گردد. پیاده‌سازی صورت گرفته از ۱۱ تابع اصلی‌ تشکیل گردیده که بر روی یک زیر مجموعه‌ی ۱۰۰تایی‌ از ۱۰۰ بازبینی ابتدایی موجود در پایگاه داده‌ی بازبینی‌ فیلم[[28]](#footnote-28) آموزش دیده و مقدار سرگشتگی و پارامتر‌های مدل را محاسبه کرده و به عنوان خروجی بر می‌‌گرداند.  
برای اجرای پیاده‌سازی صورت گرفته کافیست دستور زیر در خط فرمان نرم افزار متلب اجرا شود:

[alpha,beta,lambda,phi,sentiment\_gamma] = ldamain('train',20)

1. **منابع**

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.

[2] Lin, Chenghua, et al. "Weakly supervised joint sentiment-topic detection from text." *Knowledge and Data Engineering, IEEE Transactions on* 24.6 (2012): 1134-1145.

1. Latent Dirichlet Allocation (LDA) [↑](#footnote-ref-1)
2. Sentiment [↑](#footnote-ref-2)
3. Topic Model [↑](#footnote-ref-3)
4. Generative Probabilistic Models [↑](#footnote-ref-4)
5. Classification [↑](#footnote-ref-5)
6. Mixture Distribution [↑](#footnote-ref-6)
7. Multinomial [↑](#footnote-ref-7)
8. Gamma Function [↑](#footnote-ref-8)
9. Marginal Distribution [↑](#footnote-ref-9)
10. Latent Topic [↑](#footnote-ref-10)
11. Probabilistic Graphical Model [↑](#footnote-ref-11)
12. Observed [↑](#footnote-ref-12)
13. Joint Sentiment-Topic(JST) [↑](#footnote-ref-13)
14. Prior Information [↑](#footnote-ref-14)
15. Transform [↑](#footnote-ref-15)
16. Polarity [↑](#footnote-ref-16)
17. Inference [↑](#footnote-ref-17)
18. Posterior Distribution [↑](#footnote-ref-18)
19. Variational Inference I [↑](#footnote-ref-19)
20. Laplace Approximation [↑](#footnote-ref-20)
21. Variational Approximation [↑](#footnote-ref-21)
22. Monte Carlo Markov Chain(MCMC) [↑](#footnote-ref-22)
23. Jensen’s Inequality [↑](#footnote-ref-23)
24. Log LIkelihood [↑](#footnote-ref-24)
25. Empirical Bayes Method [↑](#footnote-ref-25)
26. Variational Expectation Maximization [↑](#footnote-ref-26)
27. Perplexity [↑](#footnote-ref-27)
28. Movie Review [↑](#footnote-ref-28)