تحلیل احساس در رسانه‌های اجتماعی فارسی با رویکرد شبکه عصبی پیچشی

yyyy ،xxxx ،yyyy ،xxxx ،yyyy ،xxxx

چکیده: افزایش کاربری شهروندان از رسانه‌های اجتماعی (مانند توئیتر، فروشگاه‌های برخط و غیره) آن‌ها را تبدیل به منبعی عظیم برای تحلیل و درک پدیده‌های گوناگون کرده است. هدف تحلیل احساس استفاده از داده‌های بدست آمده از این رسانه‌ها و کشف گرایش‌های پیدا و پنهان کاربران نسبت به موجودیت‌های خاص حاضر در متن است. در کار حاضر ما با استفاده از شبکه عصبی پیچشی، که نوعی شبکه عصبی پیش‌خور است، به تحلیل گرایش نظرات در رسانه‌های اجتماعی در دو و پنج سطح و با در نظر گرفتن شدت آن‌ها می‌پردازیم. در این شبکه عمل کانولوشن با استفاده از صافی‌هایی با اندازه‌های مختلف بر روی بردارهای جملات ورودی اعمال می‌شود و بردار ویژگی حاصل به‌عنوان ورودی لایه نرم بیشینه برای دسته‌بندی نهایی جملات بکار می‌رود. شبکه‌های عصبی پیچشی با پارامترهای مختلف با استفاده از معیار مساحت زیر منحنی و بر روی مجموعه داده جمع‌آوری شده از رسانه‌های اجتماعی فارسی ارزیابی شدند و نتایج بدست آمده نشان‌دهنده بهبود کارایی آن‌ها در گستره رسانه‌های اجتماعی نسبت به روش‌های سنتی یادگیری ماشین به‌خصوص بر روی داده‌ها با طول کوتاه‌تر هستند.

واژه‌های کلیدی: تحلیل احساس، رسانه‌های اجتماعی، شبکه عصبی پیچشی، شدت نظرات، متون کوتاه

Convolutional Neural Networks for Sentiment Analysis in Persian Social Media

Morteza Rohanian, MSc[1], Mostafa Salehi, PhD[2]

1- Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran, Email: rohanian@ut.ac.ir
2- Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran, Email: mostafa_salehi@ut.ac.ir

Abstract: With the social media engagement on the rise, the resulting data can be used as a rich resource for analyzing and understanding different phenomena around us. A sentiment analysis system employs these data to find the attitude of social media users towards certain entities in a given document. In this paper we propose a sentiment analysis method for Persian text using Convolutional Neural Network (CNN), a feedforward Artificial Neural Network, that categorize sentences into two and five classes (considering their intensity) by applying a layer of convolution over input data through different filters. We evaluated the method on three different datasets of Persian social media texts using Area under Curve metric. The final results show the advantage of using CNN over earlier attempts at developing traditional machine learning methods for Persian texts sentiment classification especially for short texts.

Keywords: Sentiment Analysis, Social Media, Convolutional Neural Network, Sentiment Intensity , Short Texts

تاریخ ارسال مقاله:
تاریخ اصلاح مقاله:
تاریخ پذیرش مقاله:
نام نویسنده مسئول: مصطفی صالحی
نشانی نویسنده مسئول: ایران، تهران، خیابان کارگر شمالی، بعد از پل جلال آل احمد، روبروی کوچه دهم، دانشکده علوم و فنون نوین، اتاق ۳۳۷.

۱- مقدمه

گسترش رسانه‌های اجتماعی امروز به افراد جامعه با نگاه‌های مختلف فرصت داده تا در فضای عمومی نظرات خود را درباره پدیده‌های گوناگون با هم به اشتراک بگذارند. ۶۹ درصد کاربران بالغ اینترنت از رسانه‌های اجتماعی به‌عنوان محلی برای بحث درباره موضوعات مختلف و اطلاع از نظرات دیگران استفاده می‌کنند [۱]. محتوای تولید شده از فعالیت در این رسانه‌ها، که ۲۸ درصد از حضور برخط کاربران را شامل می‌شود [۲]، می‌تواند منبعی عظیم داده برای تحلیل و درک رفتار افراد در مواجهه با پدیده‌های مختلف باشد.

تحلیل احساس شاخه‌ای از پردازش زبان طبیعی است که به تحلیل گرایش‌های مردم نسبت به موجودیت‌های خاص و ویژگی‌های مرتبط به آن‌ها به‌صورت خودکار می‌پردازد. هدف تحلیل احساس متمرکز بر نظراتی در زبان طبیعی است که به صورت مشخص یا ضمنی دارای جهت گیری منفی، خنثی یا مثبت هستند. این نظرات می‌توانند دارای ۵ مرتبه از نظر شدت جهت‌گیری باشند: مثبت احساسی، مثبت منطقی، خنثی، منفی منطقی و منفی احساسی [۳].

پژوهش‌های زیادی روش‌های معمول در یادگیری ماشین را در امر تحلیل احساس مورد بررسی قرار داده‌اند [۴]. رویکردهای معمول در این پژوهش‌ها اغلب بر پایه الگوریتم‌های بانظارت[1] و ویژگی‌های استخراج شده به‌صورت غیر خودکار بوده است [۵]-[۷]. این گزینش ویژگی به صورت دستی انجام‌می‌شود و بسته به موضوع و نوع متن متفاوت است. به همین دلیل مدل‌ها وابسته به متن بوده و حالت کلی و عمومی[2] ندارد. روش‌های یادگیری عمیق در سال‌های اخیر به عنوان مجموعه روش‌هایی با تعمیم پذیری بالا مورد توجه پژوهشگران حوزه پردازش زبان طبیعی بوده است و استفاده از آن در تحلیل احساس به‌خصوص برای زبان انگلیسی نیز رایج شده است. لازمه‌ی استفاده از مدل‌های یادگیری عمیق، داشتن داده آموزشی کافی، زمان و منابع رایانشی مناسب برای آموزش درست مدل شبکه عصبی است. امروزه الگوریتم‌های سنتی یادگیری ماشینی به مرور جای خود را به روش های یادگیری عمیق در تحلیل احساس می دهند. دلیل آن این است که این روش‌ها امکان این را دارند که بدون دخالت انسانی، ویژگی‌های پیچیده فراوانی درباره داده را استخراج کنند [۸]-[۱۰].

در کار حاضر برای تحلیل احساس متن فارسی ما از شبکه‌های عصبی پیچشی[3] (CNN) که نوعی شبکه عصبی پیش‌خور[4] و چند لایه هستند استفاده می‌کنیم که برای بدست آوردن خروجی به جای اتصال هر نورون لایه ورودی به لایه خروجی، بر روی داده ورودی (بردارهای کلمات حاصل از جاسازی کلمات[5]) با استفاده از صافی‌های مختلف عمل کانولوشن صورت می‌گیرد. از آن‌جا که شبکه‌های پیچشی توان استخراج ویژگی از واحدهای زبانی با طول‌های مختلف را دارند استفاده از آن‌ها نتایج بهتری نسبت به روش‌های سنتی یادگیری ماشین برای زبان‌ها با منابع فراوان به‌بار می‌آورد [۱۱]. در این پژوهش شبکه‌های عصبی پیچشی برای اولین بار برای تحلیل احساس زبان فارسی بر روی داده‌های جمع‌آوری شده از اخبار و توئیتر بکار رفته است. داده‌ها دارای دو نوع برچسب‌گذاری دو و پنج‌تایی هستند و دارای کاربری و طول متغیرند. نتایج بدست آمده در این مقاله نشان می‌دهد که این شبکه‌ها برای زبان‌ها با منابع محدود مثل فارسی هم کارآیی بهتری نسبت به روش‌های سنتی یادگیری ماشین دارند. در دسته‌بندی داده‌های متنی با طول کم این بهبود نتایج تا ۱۲ درصد رسیده است. مهم‌ترین نوآوری‌های ما در این مقاله به‌صورت زیر است:

- استفاده از شبکه‌های عصبی پیچشی برای دسته‌بندی جملات در متن فارسی و مقایسه کارآیی سه مدل با پارامترهای متفاوت از آن‌ها با روش‌های سنتی یادگیری ماشین
- تحلیل احساس بر روی واحدهای زبانی با طول‌های متفاوت و بررسی روش پیشنهادی بر روی گستره‌ای از متون فارسی با سبک‌ها و کاربری‌های متفاوت
- تحلیل احساس در ۵ سطح مختلف برای زبان فارسی، در نظر گرفتن شدت قطبیت و همین‌طور شناسایی جملات با بار خنثی
- تهیه مجموعه داده تحلیل احساس توئیتر و اخبار فارسی

در ادامه کار حاضر ما در بخش ۲ به پژوهش‌های مرتبط با کار تحلیل احساس و در بخش ۳ به معرفی روش خود می‌پردازیم. در بخش ۴ نتایج حاصل از روش پیشنهادی را گزارش می‌کنیم و درباره آن‌ها بحث می‌کنیم و بخش ۵ به نتیجه‌گیری اصلی و نمای کارهای آینده اختصاص دارد.

۲- کارهای مرتبط

بیشتر مطالعات قبلی بر اساس الگوریتم‌های یادگیری بانظارت انجام گرفته است. با این حال این الگوریتم‌ها نیاز به تهیه داده برچسب خورده دارد که به راحتی در دسترس نیست. مدل بیز ساده[6]، ساده‌ترین و پراستفاده‌ترین الگوریتم احتمالاتی برای دسته‌بندی است و بر مبنای قضیه بیز قرار دارد. این مدل احتمالات پسین رویدادها را محاسبه کرده و برچسبی که بیشترین احتمال پسین را دارد به رویداد نسبت می‌دهد.

[1] Supervised Learning
[2] Generic
[3] Convolutional Neural Networks
[4] Feedforward
[5] Word Embedding
[6] Naïve Bayes classifier

دیگردسته‌بندی کننده پرکاربرد آنتروپی بیشینه[7]، مدل احتمالاتی است که می‌توان کار دسته‌بندی را با آن انجام داد. این روش بر پایه مدل نمایی[8] و اصل حداکثرآنتروپی[9] است [۱۲]. استفاده ازاین روش تجربه-های موفقی در کار پردازش زبان طبیعی از جمله در تحلیل احساس به ارمغان آورده‌است [۱۳]. این روش در اکثر (و نه در همه) مواقع نسبت به مدل بیز ساده برتری دارد [۴]. ماشین بردار پشتیبانی[10] (SVM) برای کار دسته‌بندی اسناد بر مبنای موضوعات مشابه بسیار مفید است [۱۴]. روش SVM یک مدل یادگیری ماشینی بانظارت است که داده‌های ورودی را بررسی کرده و از میان آن‌ها الگوهایی را استخراج می‌کند که از آن‌ها می‌توان برای دسته‌بندی استفاده‌کرد. برتری این روش نسبت به دیگر روش‌های مطرح یادگیری ماشین آن است که در اینجا مدل در مورد داده‌های ورودی پیش‌فرضی ندارد و به جای تکیه بر ارزش‌های احتمالاتی، سعی دارد تا بهینه‌ترین دسته‌بندی را با داده‌های موجود انجام‌دهد. روش SVM یکی از پراستفاده‌ترین و پربازده‌ترین روش‌های یادگیری ماشینی است که در کار تحلیل احساس استفاده شده است و نتایج بدست آمده از آن برتری محسوسی به دیگر روش‌های یادگیری ماشین دارد [۶].

در سال‌های اخیر روش‌های یادگیری عمیق به‌خصوص شبکه‌های عصبی بازگشتی[11] (RNN) در تحلیل احساس برای زبان انگلیسی [۱۵]، چینی [۱۶] و آلمانی [۱۷] در میان زبان‌های مختلف با استفاده از بردارهای مختلف نمایش کلمات کاربرد زیادی داشته است. آن‌ها برای درک و کنترل ترکیب معنایی در کارهای پیچیده‌ای مانند تحلیل احساس مفید نشان داده‌شده‌اند. شبکه‌های RNN برای داده‌هایی با قابلیت تبدیل به مقادیر متوالی به کار می‌روند و با استفاده از ایده‌ی اشتراک گذاری پارامترها برای رسیدن به وزن‌های مطلوب توانایی پردازش توالی‌هایی با طول‌های متفاوت را دارند [۱۸]. با وجود این‌که استفاده از آن‌ها در تحلیل احساس برای زبان انگلیسی با نتایجی بهتر از روش‌های یادگیری بانظارت همراه بوده است [۹] به دلیل عدم اطمینان از نحوه رفتار ماتریس‌ها در مرحله بازپخش، این مدل‌ها نیاز به ویرایش دارند.

شبکه‌های پیچشی که کولوبرت و دیگران [۱۹] در ابتدا برای کاربرد در بینایی رایانه‌ای ارائه کرده‌اند اخیرا در بسیاری از کارهای پردازش زبان طبیعی مانند برچسب زنی اجزای کلام، تجزیه نحوی، تجزیه سطحی، برچسب زنی نقش معنایی[12]، و قطعه بندی[13] مورد استفاده قرار گرفته‌است. استفاده از شبکه‌های پیچشی در تحلیل احساس نیز برای زبان‌ها با منابع فراوان مورد استفاده قرار گرفته و باعث بهبود قابل توجه دقت و کاهش زمان مرحله آموزش نسبت به دیگر روش‌های یادگیری عمیق شده است [۱۰].

پژوهش‌های حوزه تحلیل احساس در زبان فارسی معمولا یا با استفاده از روش‌های مبتنی بر قاعده هستند یا مبتنی بر پیکره [۲۴]. برای بهبود نتایج معمولا از پیش پردازش نظرات و ویژگی‌های لغت‌نامه استفاده شده است [۲۵]. بازدهی این روش‌ها وابسته به کیفیت برچسب‌دهی در پیکره‌ها و شیوه گزینش ویژگی‌ها پیش از شروع کار دسته‌بندی است.

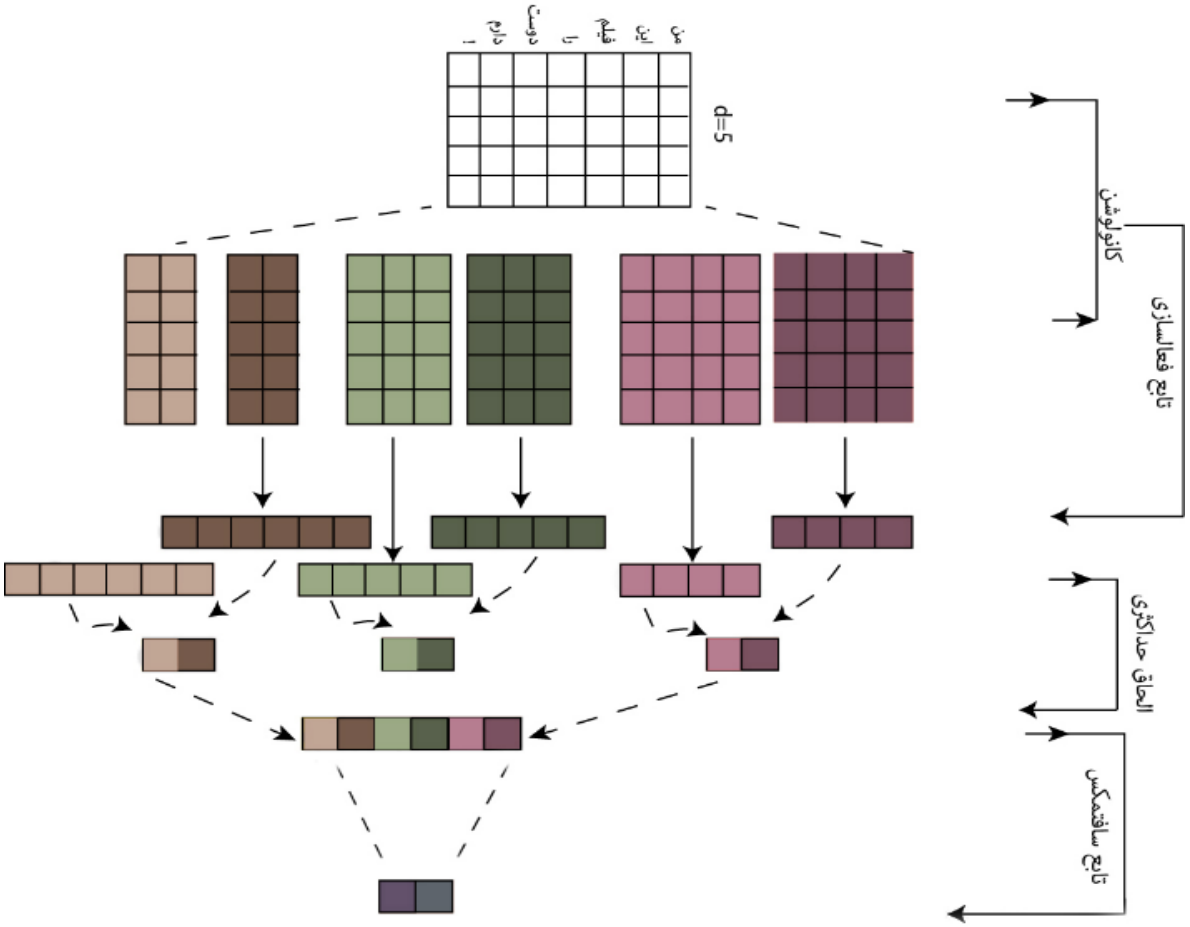
به‌طور کلی مزایای استفاده از یادگیری عمیق شامل موارد زیر است [۲۰]:

- احتیاجی نیست ویژگی‌ها به صورت دستی تهیه شوند. در یادگیری عمیق به جای استخراج دستی ویژگی‌ها از جاسازی کلمات استفاده می‌شود که در آن‌ها اطلاعات مربوط به بافت متنی وجود دارد ودر مرحله آموزش، لایه‌های میانی شبکه عصبی به صورت خودکار ویژگی‌ها را استخراج می‌کنند.
- با استفاده از شبکه‌های عصبی یادگیری، انتخاب ویژگی‌ها[14] و نمایش آن‌ها می‌تواند هم با یادگیری بانظارت و هم بدون نظارت[15]صورت گیرد.
- در تحلیل احساس با متن‌های گوناگونی از لحاظ سبک نوشتار و بافت معنایی روبرو هستیم. انعطاف و تعمیم‌پذیری روشهای یادگیری عمیق، اجازه می‌دهد تا با مشکل عدم تعمیم پذیری مدل کم‌تر روبرو شویم.

۳- راهکار پیشنهادی

برای تحلیل احساس زبان‌ها با منابع محدود مثل فارسی با استفاده از یادگیری عمیق نیاز به مجموعه‌ای از بردارها برای نمایش کلمات داریم که آن‌ها را با استفاده از جاسازی کلمات روی مجموعه ویکی‌پدیای فارسی بدست می‌آوریم. این بردارها به‌عنوان داده ورودی شبکه برای استخراج ویژگی‌ها به‌کار می‌روند. شبکه‌های عصبی پیچشی (CNN) نوعی خاص از شبکه‌های عصبی برای پردازش داده هستند که با این بردارهای کلمات مانند یک تور[16] برخورد می‌کنند. صافی‌های[17] هر لایه کانولوشن بر روی طول ماتریس‌های حاصل از بردارهای ورودی حرکت می‌کند. عرض صافی‌ها به اندازه عرض بردار ورودی (بعد بردار کلمات) و طول آن‌ها معمولا بین ۲ تا ۵ کلمه است. از بردارهای حاصل نگاشت‌های ویژگی[18] حاصل می‌شوند که با استفاده از لایه الحاق حداکثری[19] تبدیل

[7] Maximum Entropy Classifier
[8] Exponential Model
[9] Principle of Maximum Entropy
[10] Support Vector Machines (SVMs)
[11] Recurrent Neural Networks
[12] Semantic Role Labeling
[13] Chunking
[14] Representation Learning
[15] Unsupervised Learning
[16] Grid
[17] Filters
[18] Feature Maps
[19] Max-pooling



شکل ۱ – نمایش یک شبکه عصبی پیچشی برای دسته‌بندی جملات.

به یک بردار نهایی می‌شوند که از آن به‌عنوان ورودی لایه آخر برای دسته‌بندی جملات استفاده می‌شود.

پایه اصلی شبکه پیچشی بردارهای کلمات ورودی $x \in R^s$ هستند که در آن s ابعاد بردارهاست و هر سند ورودی به‌صورت ماتریس $d \in R^{n\times s}$ نمایش داده می‌شود که n تعداد کلمات آن است و هر ردیف ماتریس، بردار یک کلمه را نمایندگی می‌کند.

لایه کانولوشن که هدف از آن استخراج ویژگی‌های محلی از رشته‌های حرفی داخل جمله است، تعمیمی از رویکرد پنجره است که در آن چند صافی با اندازه مشخص کل جمله را می‌پیمایند و نتایج حاصل از کانولوشن روی ماتریس‌های مختلف را با هم ترکیب می‌کنند. این صافی‌ها عرضی به اندازه ماتریس ورودی و طولی قابل تعیین دارند. این صافی‌ها همان‌طور که در شکل ۱ مشخص است خود هر کدام یک ماتریس‌اند که مجموع حاصل‌ضرب عناصر آن و ماتریس ورودی خود ماتریس جدیدیست که بردار کلمه ورودی را با توجه به بافتش تغییر می‌دهد. ما یک لایه کانولوشن $W \in R^{h\times s}$ تعریف می‌کنیم که در آن h تعداد کلماتی است که ما می‌خواهیم روی آن کانولوشن انجام دهیم (طول صافی). اگر کانولوشن را با عمل‌گر × نشان دهیم هر بار پیمایش صافی روی بردار برابر است با:

$$W * d_{j:j+h-1} = \sum_{i=j}^{j+h-1}\sum_{k=0}^{s-1} W_{i,k} d_{i,k}. \quad (۱)$$

سپس صافی به طول h کلمه، $d_{j:j+h-1}$، را با استفاده از تابع غیرخطی f به عدد حقیقی c_j نگاشت می‌کنیم:

$$c_j = f(d_{j:j+h-1} + b), \quad (۲)$$

که b در آن یک عدد حقیقی است که میزان تمایل[20] را نشان می‌دهد. با اعمال کانولوشن بر روی تمام سند با استفاده از W، یک بردار ویژگی حاصل می‌شود

$$c(W) = [c_1, c_2, \dots, c_{n-h+1}]. \quad (۳)$$

هرشبکه می‌تواند دارای مقادیر متفاوت برای نوع صافی‌ها و ماتریس‌های وزنی باشد که به هر کدام کانال[21] گفته می‌شود. در حین مرحله آموزش یک شبکه عصبی پیچشی بر پایه کاربری خاص صورت گرفته، عناصر ماتریس‌های صافی‌ها یاد گرفته می‌شوند. چون هر عنصر ورودی و صافی

[20] Bias

[21] Channel

باید به طور جداگانه ذخیره‌سازی شوند، معمولا فرض می‌کنیم که این عناصر جز در نقاط محدودی که مقادیر ذخیره شده‌اند، دارای مقدار صفر هستند. طول خروجی در این لایه به طول جمله ورودی و تعداد کانال-های استفاده شده بستگی دارد.

هدف از لایه الحاق حداکثری، یکسان‌سازی طول بردارهای جملات[۲۲] و کاهش ابعاد بردار خروجی در عین حفظ اطلاعات مهم است. برای مثال اگر ۵۰۰ صافی وجود داشته باشد بعد از اعمال الحاق حداکثری ما برداری ۵۰۰ بعدی داریم که به‌عنوان ورودی با طول ثابت سافتمکس استفاده می‌شود. هم‌چنین اگر هر صافی را دارای اطلاعات مربوط به یک ویژگی خاص ورودی بدانیم با استفاده از الحاق حداکثری می‌توانیم پیش‌بینی کنیم که آیا این ویژگی در جمله وجود داشته یا نه. کار الحاق حداکثری در این لایه این است که بیشترین مقدار هر ویژگی را از میان صافی‌های مختلف برگزیند

$$\hat{c}_W = \max c(W)_i. \qquad (۴)$$

این روش نسبت به میانگین‌گیری بهتر است چرا که در دسته‌بندی، همه کلمات به یک اندازه مهم نیستند و اهمیت نسبی آن‌ها در لایه الحاق حداکثری لحاظ می‌شود. در نهایت از همه صافی‌ها یک بردار ویژگی سراسری[۲۳] بدست می‌آید که ورودی لایه بعد محسوب می‌شود:

$$\hat{C}_W = [\hat{c}_{W^1}, \dots, \hat{c}_{W^k}]^T, \qquad (۵)$$

که در آن $T = \{1, \dots, k\}$ است. اندازه بردار ویژگی سراسری برای جملات مختلف ثابت است و اندازه آن به کلمات جداگانه و تعداد کانال‌ها وابسته است. در این مرحله اطلاعات مربوط به مکان قرار گیری ویژگی‌های مختلف را با توجه به در نظر نگرفتن ترتیب کلمات از دست می‌دهیم.

در لایه آخر از تابع نرم بیشینه[۲۴]، که نوع گسترش داده شده رگرسیون لجستیک است، برای دسته‌بندی بردار ویژگی بدست آمده استفاده می‌کنیم. اگر $U \in R^{2\times k}$ و $b^U \in R^2$ پارامترهای لایه نرم بیشینه باشند و ورودی وزن‌دار برابر باشد با:

$$y_j = U_j \hat{c}_W + b_j^U \qquad (۶)$$

که در آن c_w بردار ورودی، U_j ردیف j–ام U و b_j^U عنصر j–ام b^U و Y_j برچسب j–ام در ماتریس d است. اندازه این لایه برابر با تعداد برچسب‌ها است. احتمال برچسب خروجی برابر است با:

$$P(Y_j = 1 | d, W, b, U, b^U) = \frac{e^{y_j}}{\sum_i e^{y_i}}, \qquad (۷)$$

اگر D مجموعه ماتریس‌های داده آموزش باشد برای آموزش دسته‌بندی دوتایی باید فرمول‌های زیر را به ترتیب برای داده‌های منفی و مثبت کمینه کرد:

$$-\sum_{d\in D} \log(P(Y_{POS}^d - 1 | d, W, b, U, b^U)) \qquad (۸)$$

$$-\sum_{d\in D} \log(P(Y_{POS}^d - 2 | d, W, b, U, b^U)) \qquad (۹)$$

پارامترهای شبکه عصبی پیچشی (d, W, b, U, b^U) که فرمول بالا را کمینه می‌کنند با محاسبه گرادیان از طریق روش پس‌انتشار بدست می‌آیند. شکل کلی معماری CNN پیشنهادی در شکل ۲ قابل مشاهده است.

۳-۱- جاسازی کلمات

شبکه‌های عصبی تنها اعداد و توابع را به رسمیت می‌شناسند. بنابراین بر خلاف روش‌های دیگر در یادگیری ماشین، ورودی در اینجا به صورت متنی نیست. این به آن معناست که برای تهیه ورودی، متن باید به بردارهای ویژگی یا به عبارت دیگر به جاسازی‌های کلمات تبدیل شوند که این جاسازی‌ها، اگر درست استخراج بشوند خود حاوی اطلاعات بافتی و معنایی متن هستند [۲۱].

این بردارها با آموزش دادن شبکه عصبی بر مبنای پیکره متنی حاصل می‌شوند و فرایندی زمان‌بر هستند با استفاده از یادگیری عمیق، یک مدل زبانی برای یادگیری *نمایش توزیع یافته کلمات*[۲۵] بر اساس سه ایده کلی زیر می‌توان ارائه‌کرد [۲۲]:

آ. هر کلمه در پیکره به یک بردار ویژگی s بعدی حاوی اعداد حقیقی متناظر می‌شود.

ب. تابع احتمال توام برای کلمات، با استفاده از این نمایش‌های برداری بیان می‌شود.

ج. یادگیری بردارهای ویژگی کلمات و پارامترهای تابع احتمال به طور همزمان انجام می‌شوند.

هر کدام از جاسازی‌های کلمات می‌تواند ابعادی به دلخواه کاربر داشته-باشد. بعد بالاتر به معنای این است که اطلاعات بیشتری ضبط شده است ولی در عین حال با زیاد شدن بعد هزینه‌های محاسباتی نیز افزایش‌می-یابد.

۴- ارزیابی راهکار پیشنهادی

برای ارزیابی راهکار پیشنهادی و مقایسه‌ی نتایج آن با روش‌های سنتی یادگیری ماشین، آن را با پارامترهای مختلف بر روی مجموعه داده‌ی رسانه‌های اجتماعی فارسی آزمایش می‌کنیم.

۴-۱- مجموعه داده

مدل ما در کار حاضر بر روی داده ۲.۵ گیگابایت ویکی‌پدیای فارسی با آموزش دیده شده است [۲۳]. برای ارزیابی مدل از مجموعه داده‌های مختلف استفاده شده است:

- مجموعه داده سنتی‌پرس: مجموعه ۱۱۰۰ نظر درباره محصولات دیجیتال که از فروشگاه برخط دیجی‌کالا جمع-آوری شده است.[۲۶].

[22] Sentence Vector
[23] Global Feature Vector
[24] Softmax
[25] Distributed Representation of Words

- مجموعه داده توئیتر فارسی: مجموعه ۱۵۰۰۰ توئیت فارسی جمع‌آوری شده برای ارزیابی و آموزش در بازه بین فروردین تا تیر ۱۳۹۶ .
- مجموعه داده اخبار: ۴۳۰۰ خبر فارسی جمع‌آوری شده از ۳۰ منبع خبر فارسی در بازه بین فروردین تا تیر ۱۳۹۶.

۱۰ درصد هر کدام از مجموعه داده‌ها به‌عنوان داده ارزیابی و ۹۰ درصد بقیه داده آموزش در نظر گرفته شده است.

۴-۲- پارامترهای انتخابی

برای آموزش همه مجموعه داده‌ها از پنجره‌هایی صافی‌هایی سه‌تایی و ابعاد ۲۰۰ برای بردارهای ورودی استفاده شده است. در مرحله آموزش تعداد تکرارهای ۱۰۰ بکار رفته است. تعداد دسته‌ها دو و پنج در نظر گرفته شده و برای تنظیم کردن از دراپ‌اوت با نرخ ۰.۵ در لایه ماقبل آخر استفاده شده است. آموزش با استفاده از گرادیان کاهشی اتفاقی و قانون به‌روز رسانی آدادلتا صورت گرفته است. برای رسیدن به تحلیل مناسب شرایط تصادفی (مقداردهی اولیه برای شبکه و جاسازی کلمات و...) برای همه مدل‌ها یکسان در نظر گرفته شده است.

۴-۳- معیار ارزیابی

برای ارزیابی از مساحت سطح زیر نمودار دو بعدی که در آن نرخ تشخیص صحیح دسته مثبت روی محور Y و نرخ تشخیص غلط دسته منفی روی محور X رسم می‌شود استفاده می‌کنیم که به آن مساحت زیر منحنی[۲۶] (AUC) می‌گوییم. هرچه عدد زیر نمودار بزرگ‌تر باشد دسته‌بندی صورت گرفته دقیق‌تر بوده است.

۴-۴- نتایج

نتایج مدل‌های ما با پارامترهای متفاوت برای دو و پنج دسته در مقایسه با مدل‌های یادگیری ماشین سنتی در جدول ۱ و ۲ آمده است. ارزیابی‌ها با معیار مساحت زیر نمودار نشان می‌دهند که مدل شبکه‌های پیچشی به‌طور قابل ملاحظه‌ای عملکرد بهتری در هر دو دسته نسبت به روش‌های پراستفاده سنتی یادگیری ماشین دارند. تعدد لایه‌ها برای پردازش، که ویژگی‌های سطح بالا را برخلاف روش‌های سنتی یادگیری ماشین بدون نظارت استخراج می‌کنند و وجود لایه الحاق حداکثری که نمونه‌های مناسبی از ویژگی‌ها را انتخاب می‌کند به بهبود نتایج کمک کرده‌اند. با وجود این‌که برخلاف بسیاری روش‌های دیگر یادگیری ماشین شبکه عصبی پیچشی ترتیب کلمات را محسوب نمی‌کند، صافی‌های موجود شبکه تا پنج کلمه و بیشتر را در کنار هم در نظر می‌گیرند که محاسبه آن در روش‌های سنتی بسیار دور از ذهن است.

این موضوع به‌خصوص در جملاتی که در آن‌ها، بین بخشی از عبارتی خاص با قرار گرفتن کلمات دیگر فاصله افتاده بکار می‌آید.
به‌طور کلی افزایش بعد و تعداد تکرار در مرحله آموزش بردارهای ورودی کلمات به بهتر شدن کارآیی مدل منجر می‌شود. این بردارها که از متون ویکی‌پدیای فارسی بدست آمده‌اند برای کاربری ویژه تحلیل احساس دیگر روابط معنایی موجود در بردارهای کلمات است، که محصول نمایش توزیع یافته آن‌هاست و توانایی شبکه عصبی پیچشی را برای کشف روابط معنایی ترکیبی اجزای جمله نسبت به روش‌های دیگر یادگیری ماشین افزایش می‌دهد.

جدول ۱- نتایج ارزیابی مدل‌ها برای دسته‌بندی ۲ و ۵تایی روی داده‌های مختلف با معیار مساحت زیر نمودار

	پنج‌تایی			دوتایی		
مدل	توئیتر	اخبار	نظرات	توئیتر	اخبار	نظرات
CNN	۴۳.۸	۴۶.۱	۴۳.۹	۷۵.۳	۷۹.۸	۷۳.۱
NB	۳۹.۷	۴۰.۲	۳۸.۳	۵۲	۶۲.۳	۵۶
MEC	۳۹.۱	۳۹.۸	۳۸	۶۱.۶	۷۱	۶۸.۳
SVM	۴۰.۱	۴۰.۴	۴۰.۱	۶۳.۷	۷۴.۱	۶۸.۱

هم‌چنین نتایج بدست آمده نشان‌دهنده کارایی بالاتر شبکه‌های عصبی پیچشی بر روی رسانه‌های اجتماعی با طول متن کوتاه‌تر (توئیتر) نسبت به روش‌های سنتی یادگیری ماشین است. اگرچه عمل‌کرد مدل پیشنهادی در محیط زبان غیررسمی و با علائم نگارشی و غیر نگارشی مرسوم با افت روبرو می‌شود. در شبکه‌های اجتماعی در تحلیل متون خبر بردارهای عمومی بدست آمده از ویکی‌پدیای فارسی، به‌دلیل استفاده از زبان رسمی، کارایی بهتری نسبت به دو داده دیگر نشان داده‌اند.

بدیهی است که علاوه بر بردارهای کلمات عمومی استفاده از بردارهای کارویژه (مخصوص تحلیل احساس در اینجا) به افزایش دقت مدل کمک خواهد کرد. همین‌طور در کار حاضر برای بدست‌آوردن بردارهای جملات ما از تجمیع بردارهای کلمات استفاده کردیم. برای بهبود کیفیت نمایش برداری متن می‌توان از روش‌های دیگری چون بردارهای پاراگراف، که قادر است متون با طول‌های مختلف را نمایندگی کند بهره گرفت.

[26] Area Under Curve

۴-۵- تاثیر تعداد صافی‌ها بر نتایج

تاثیر اندازه صافی را بر روی داده‌های مختلف و با تنظیم بعد بردارها بر روی ۲۰۰ و تعداد تکرار بر روی ۵۰ اندازه گرفته شده‌است. اندازه‌های صافی را برابر دو، سه، پنج و هفت می‌گیریم. نتایج برای دو و پنج دسته در جدول ۲ آمده است.

جدول ۲- نتایج ارزیابی مدل‌ها برای دسته‌بندی ۲ و ۵تایی روی داده‌های مختلف با در نظر گرفتن تاثیر تعداد صافی‌ها

	پنج‌تایی			دوتایی		
تعداد صافی	نظرات	توئیتر	اخبار	توئیتر	اخبار	نظرات
دو	۴۰.۷	۴۸.۷	۴۴.۴	۷۴.۳	۶۸.۴	۶۸.۸
سه	۴۳.۱	۴۷.۲	۴۶.۶	۷۵.۸	۷۸.۸	۷۰.۷
پنج	۴۲.۹	۴۶.۳	۴۵.۷	۶۹.۴	۸۰.۲	۷۲.۱
هفت	۴۵.۸	۴۲.۱	۴۴.۶	۶۳.۳	۷۹.۷	۷۳.۴

نتایج بدست آمده نشان می‌دهد که داده با طول کوتاه‌تر با صافی‌ها با اندازه کوچک‌تر بهتر عمل می‌کنند. با افزایش طول جملات نیاز به صافی‌های بزرگ‌تر احساس می‌شود. عملکرد مدل برای دسته‌بندی دوتایی توئیتر با افزایش اندازه فیلترها به‌شکل محسوسی افت و برای دسته‌بندی نظرات و اخبار با بزرگ شدن فیلتر در هر دو بخش دو و پنج‌تایی رشد می‌کند. مناسب‌ترین اندازه صافی برای دسته‌بندی متون توئیتری در هر دو بخش سه مشاهده می‌شود. مشاهدات ما نشان‌دهنده این است که بهتر است پیش از آموزش مدل صافی با اندازه مناسب از راه آزمون و خطا بدست آید. در هر دل هر صافی را می‌توان به تنهایی یا به‌صورت ترکیبی با صافی‌های دارای اندازه نزدیک بکار برد.

۴-۶- تاثیر ابعاد بردارهای کلمات بر نتایج

با قرار دادن اندازه صافی روی پنج و تعداد تکرار ۵۰، ما ابعاد را برابر اعداد ۱۰، ۵۰، ۱۰۰، ۲۰۰ و ۳۰۰ قرار دادیم. بهترین عدد برای بعدانتخابی به نوع مجموعه داده بستگی دارد.

از نتایج نشان‌داده شده در جدول ۳ مشاهده می‌شود که عملکرد مدل با افزایش ابعاد بردارها به بیش از ۱۰ بهبود چشم‌گیری پیدا می‌کند. این موضوع بر رود داده اخبار که بیشترین طول را دارد محسوس‌تر است. می‌توان نتیجه گرفت افزایش ابعاد به بیش از ۲۰۰ در کار حاضر تاثیر چندانی در نتایج ندارد و حتی ممکن است کارآیی را کاهش دهد. همین‌طور با افزایش ابعاد بردارهای کلمات زمان مورد نیاز برای آموزش آن‌ها به طور قابل توجهی افزایش می‌یابد. در کاربرد به نظر می‌رسد تعداد دسته‌ها و ابعاد بردار رابطه قابل مشاهده‌ای ندارند و برای انواع دسته‌بندی‌ها بازه ۱۰۰ تا ۲۰۰ مناسب است.

جدول ۳- نتایج ارزیابی مدل‌ها برای دسته‌بندی ۲ و ۵تایی روی داده‌های مختلف با در نظر گرفتن تاثیر ابعاد بردارهای کلمات

	پنج‌تایی			دوتایی		
ابعاد بردارها	نظرات	توئیتر	اخبار	توئیتر	اخبار	نظرات
۱۰	۴۱.۷	۴۰.۳	۳۳.۶	۶۱.۶	۶۶.۶	۶۱.۳
۵۰	۴۳.۱	۴۵.۲	۴۳.۱	۷۵.۲	۷۸.۸	۷۲.۷
۱۰۰	۴۳.۹	۴۶.۳	۴۴.۱	۷۶.۴	۷۹.۲	۷۴.۴
۲۰۰	۴۲.۸	۴۶.۱	۴۳.۹	۷۵.۳	۷۹.۱	۷۳.۵
۳۰۰	۴۲	۴۵.۷	۴۳.۳	۷۴.۹	۷۸.۵	۷۲.۳

پیدا کردن بازه مناسب برای داده‌های آموزش با اندازه بزرگ‌تر از داده حاضر نیاز به بررسی بازه‌های متفاوت دارد.

۴-۵- تاثیر تعداد تکرار در آموزش بردار کلمات بر نتایج

برای بررسی عدد مناسب تکرار در مرحله آموزش بردارهای کلمات ما ابعاد بردار کلمات را ۱۰۰ و اندازه صافی را پنج در نظر می‌گیریم. اعداد تکرار ما ۲۵، ۵۰، ۱۰۰ و ۲۰۰ است.

جدول ۴- نتایج ارزیابی مدل‌ها برای دسته‌بندی ۲ و ۵تایی روی داده‌های مختلف با در نظر گرفتن تاثیر تعداد تکرار در آموزش بردار کلمات

	پنج‌تایی			دوتایی		
تکرار آموزش	نظرات	توئیتر	اخبار	توئیتر	اخبار	نظرات
۲۵	۳۹.۷	۴۰.۷	۴۲.۶	۷۱.۸	۷۲.۵	۶۹.۷
۵۰	۴۳.۱	۴۵.۲	۴۳.۱	۷۵.۲	۷۸.۸	۷۲.۷
۱۰۰	۴۲.۹	۴۶.۳	۴۴.۱	۷۴.۴	۷۹.۲	۷۱.۱
۲۰۰	۴۴.۸	۴۵.۷	۴۲.۵	۷۵.۳	۷۷.۸	۷۱.۱

از نتایج نمایش داده شده در جدول ۴ مشاهده می‌شود که تعداد تکرار بیش‌تر از ۲۵ ارتباط مشخصی با انواع داده با طول‌های مختلف ندارد. همین‌طور نتایج دسته‌بندی‌های مختلف با تعداد دسته مختلف ارتباط معناداری با تعداد تکرار در آموزش بردارها ندارد. تکرار بیش‌تر از ۱۰۰

نتایج را تغییر چندانی نمی‌دهد. بدیهی است که با ثابت نگاه داشتن افزایش تکرار زمان طی شده برای آموزش بردارها افزایش می‌یابد. تکرار ۱۰۰ نتایج بهتری در مجموع دارد اما با توجه به فاصله بسیار نزدیکی که با تعداد تکرار ۵۰ دارد و با توجه به مدت زمان طی شده برای آموزش برای داده‌ای به حجم مجموعه ما چندان به‌صرفه به‌نظر نمی‌رسد. اصولا برای زبان‌ها با منابع محدود تعداد تکرار بین ۵۰ تا ۱۰۰ پیشنهاد می-شود.

۵- نتیجه‌گیری و پیشنهادها

در کار حاضر ما شبکه عصبی پیچشی با پارامترهای متفاوت را با استفاده از بردارهای کلمات بر روی رسانه‌های اجتماعی جهت تحلیل احساس متن فارسی به‌کار بردیم. بدون آموزش بردارهای خاص تحلیل احساس شبکه عصبی با یک لایه کانولوشن کارایی بهتری نسبت به روش‌های سنتی یادگیری ماشین به‌خصوص بر روی داده‌ها با طول کوتاه نشان داد. نتایج ما نشان داد که بردارهای کلمات استخراج شده از داده‌های عمومی بدون توجه به نوع کاربری می‌توانند به به بهبود نتایج در پردازش زبان طبیعی کمک کنند. برای تحقیقات آتی، در نظر گرفتن بردارهای کارویژه، بردارهای واحدهای متنی فراتر از کلمه و ترتیب کلمات پیشنهاد می‌شود.

مراجع

[1] Greenwood, S., A. Perrin, and M. Duggan. "Social media update 2016: Facebook usage and engagement is on the rise, while adoption of other platforms holds steady." *Pew Research Center* (2016).

[2] Mander, Jason. "Daily time spent on social networks rises to 1.72 hours." *London: Global Web Index* (2015).

[3] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5, no. 1 (2012): 1-167.

[4] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79-86. Association for Computational Linguistics, 2002.

[5] Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath. "Classification of sentiment reviews using n-gram machine learning approach." *Expert Systems with Applications* 57 (2016): 117-126.

[6] Mullen, Tony, and Nigel Collier. "Sentiment Analysis using Support Vector Machines with Diverse Information Sources." In *EMNLP*, vol. 4, pp. 412-418. 2004.

[7] Agarwal, Basant, and Namita Mittal. "Machine learning approach for sentiment analysis." In *Prominent Feature Extraction for Sentiment Analysis*, pp. 21-45. Springer International Publishing, 2016.

[8] Poria, Soujanya, Haiyun Peng, Amir Hussain, Newton Howard, and Erik Cambria. "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis." *Neurocomputing* (2017).

[9] Dong, Li, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. "Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification." In *ACL (2)*, pp. 49-54. 2014.

[10] Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

[11] Zhang, Ye, and Byron Wallace. "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification." *arXiv preprint arXiv:1510.03820* (2015).

[12] Jaynes, Edwin T. "Information theory and statistical mechanics." *Physical review* 106, no. 4 (1957): 620.

[13] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." In *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*, pp. 1-5. IEEE, 2013.

[14] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20, no. 3 (1995): 273-297.

[15] Dos Santos, Cícero Nogueira, and Maira Gatti. "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts." In *COLING*, pp. 69-78. 2014.

[16] Zhang, Yu, Mengdong Chen, Lianzhong Liu, and Yadong Wang. "An effective convolutional neural network model for Chinese sentiment analysis." In *AIP Conference Proceedings*, vol. 1836, no. 1, p. 020085. AIP Publishing, 2017.

[17] Cieliebak, Mark, Jan Deriu, Dominic Egger, and Fatih Uzdilli. "A Twitter Corpus and Benchmark Resources for German Sentiment Analysis." *SocialNLP 2017* (2017): 45.

[18] Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. "Recursive deep models for semantic compositionality over a sentiment treebank." In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, vol. 1631, p. 1642. 2013.

[19] Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. "Natural language processing (almost) from scratch." *Journal of Machine Learning Research* 12, no. Aug (2011): 2493-2537.

[20] Wang, Keze, Xiaolong Wang, Liang Lin, Meng Wang, and Wangmeng Zuo. "3D human activity recognition with reconfigurable convolutional neural networks." In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 97-106. ACM, 2014.

[21] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111-3119. 2013.

[22] Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A neural probabilistic language model." *Journal of machine learning research* 3, no. Feb (2003): 1137-1155.

[23] Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information." *arXiv preprint arXiv:1607.04606* (2016).

[24] Bagheri, Ayoub, and Mohamad Saraee. "Persian Sentiment Analyzer: A Framework based on a Novel Feature Selection Method." *International Journal of Artificial Intelligence™* 12, no. 2 (2014): 115-129.

[۲۵] حسین اکبریان و دیگران. (۱۳۹۵). تعیین جهت گیری نظرات در رسانه‌های اجتماعی فارسی زبان. ارائه شده در بیست و چهارمین کنفرانس مهندسی برق ایران، شیراز: دانشگاه شیراز

[۲۶] پدرام حسینی و دیگران. (۱۳۹۳). پیکره فارسی تحلیل احساس سنتی پرس. ارائه شده در سومین همایش ملی زبان-شناسی رایانشی، تهران: دانشگاه صنعتی شریف.