ارائه یک سیستم تحلیل احساس در زبان فارسی
با استفاده از مدل‌های یادگیری عمیق

جواد پورمصطفی
روشن شارمی
دانشجو کارشناسی ارشد
گروه مهندسی کامپیوتر
دانشکده فنی، دانشگاه گیلان
javad.pourmostafa@gmail.com

پارسا عباسی سرابستانی
دانشجو کارشناسی
گروه مهندسی کامپیوتر
دانشکده فنی، دانشگاه گیلان
parsa.abbasi1996@gmail.com

سید ابوالقاسم میرروشندل
دانشیار و عضو هیئت علمی
گروه مهندسی کامپیوتر
دانشکده فنی، دانشگاه گیلان
mirroshandel@gmail.com

چکیده

تحلیل احساس یکی از زیرشاخه‌های پردازش زبان طبیعی است که هدف آن طبقه‌بندی متون با توجه به احساس، عقیده و نگرش بیان شده در آنها است. در دهه‌های گذشته روی استفاده از رویکردهای یادگیری ماشین جهت حل مسائل تحلیل احساس کارهای زیادی انجام گرفت، اما تمرکز اصلی آنها بر روی ساخت استخراج‌کننده‌های ویژگی قوی‌تر بوده است، زیرا که عملکرد این یادگیرنده‌ها بیشتر به انتخاب نحوه بازنمایی داده وابسته است. در سال‌های اخیر با افزایش قدرت محاسباتی و پیشرفت علم یادگیری عمیق، با واگذاری یادگیری الگوها به ماشین، پیشرفت‌های چشمگیری در بسیاری از مسائل رشته‌های مختلف به‌ویژه در طبقه‌بندی متون و تحلیل احساس حاصل شده است. در تحلیل احساس، بیشتر این پیشرفت‌ها برای زبان انگلیسی صورت گرفته و در زبان فارسی به دلیل چالش‌هایی همچون عدم دسترسی به مجموعه‌داده کافی و ابزارهای دقیق پردازش متن، پیشرفت‌ها به آن میزان چشمگیر نبوده است. هدف این مقاله بررسی و مقایسه رویکردهای یادگیری ماشین و یادگیری عمیق در تحلیل احساس جملات فارسی می‌باشد. بدین منظور از بیز ساده و گرادیان کاهشی تصادفی و ماشین بردار پشتیبانی به عنوان الگوریتم‌های یادگیری ماشین و حافظه طولانی کوتاه مدت دوطرفه و شبکه عصبی پیچشی به عنوان مدل‌های یادگیری عمیق استفاده شده است. همچنین روش‌های مختلفی جهت حل چالش‌های مرتبط با مجموعه‌داده، یعنی عدم توازن و تعداد پایین اسناد ارائه و مقایسه شده است.

کلیدواژه‌ها: متن‌کاوی[1]، تحلیل احساس[2]، یادگیری عمیق[3]، یادگیری ماشین[4]، پردازش زبان طبیعی[5]

۱. مقدمه

تحلیل احساسات که همچنین نظرکاوی[6] نامیده می‌شود، بخشی از مطالعات است که به تحلیل عقاید، احساسات، سنجش‌ها و نگرش‌های مردم درباره موجودیت‌هایی همچون محصولات، سرویس‌ها، سازمان‌ها، افراد، مشکلات، رخدادها، موضوعات و خواص آنها می‌پردازد (Liu, 2012).

---

[1] Text mining
[2] Sentiment analysis
[3] Deep learning
[4] Machine learning
[5] Natural language processing
[6] Opinion mining

«تحلیل احساس» و «نظرکاوی» از نظر واژگان تفاوت ظریفی با یکدیگر دارند. یک نظر به معنی دیدگاه واقعی یک شخص درباره چیزی است اما احساسات به دریافت و احساس وی درباره آن اشاره دارد. به عنوان مثال جمله "من درباره وضعیت کنونی اقتصاد نگران هستم" یک احساس را بیان می‌کند، در حالی که جمله "من فکر می‌کنم اقتصاد خوب عمل نمی‌کند" بیانگر یک نظر است. در پاسخ به جمله نخست، می‌توان گفت "من درک‌تان می‌کنم"، اما برای جمله دوم می‌توانیم پاسخ دهیم "با شما موافق/مخالف هستم" (Liu, 2015). با این حال از آنجایی که هر دو جمله فوق می‌توانند منجر به بیان یک احساس مشابه درباره موضوع شوند، در این رشته، از هر دو اصطلاح برای مفهومی مشابه استفاده می‌شود.

رشد بسیار زیاد محتوای تولیدی توسط کاربر[7] در وب‌سایت‌ها و شبکه‌های اجتماعی مختلف همچون توییتر، فیسبوک و آمازون باعث شده این شبکه‌ها به هسته اصلی کشف عقاید درباره موضوعات مختلف تبدیل شوند. تردیدی نیست که بدون وجود این حجم از اطلاعات دیجیتالی آنلاین، امکان انجام بسیاری از تحقیقات صورت گرفته در این خصوص در سال‌های اخیر به وجود نمی‌آمد. امروزه عموم مردم دیدگاه‌های مطرح شده در شبکه‌های اجتماعی، بلاگ‌ها و توییت‌ها را دنبال می‌کنند و بسیاری از شرکت‌ها نیز به کاوش بازخوردهای محصولات و رضایت مشتریان روی آورده‌اند. از همین رو، نیازی جهت تحلیل داده‌ها به منظور تشخیص قطبیت عقاید به وجود آمده، که باعث شده چه واحدهای آموزشی و تحقیقاتی و چه صنعت بیش از پیش بدین حوزه علاقه‌مند شوند (Rojas-Barahona & Maria, 2016).

از سوی دیگر، علم یادگیری عمیق توانسته با پیشرفت خود به بسیاری از مسائل حوزه پردازش زبان طبیعی[8] پاسخ دهد و جایگزینی امیدوارکننده برای روش‌های سنتی به شمار رود. یادگیری عمیق تاکنون از خود عملکرد بسیاری خوبی در بسیاری از شاخه‌های پردازش زبان طبیعی، خصوصاً تحلیل احساسات نشان داده است. مهم‌ترین مزیت این روش، بی‌نیازی از استخراج دستی ویژگی‌ها است که به جای تخصص در حوزه زبان‌شناسی بر دسترسی به حجم بالای داده‌ها تکیه دارد.

با توجه به موارد فوق، هدف این مقاله استفاده و بررسی تکنیک‌های مختلف جهت تحلیل احساسات در زبان فارسی و در سطح جمله است. بدین منظور از مجموعه داده‌ای شامل نظرات خریداران محصولات دیجیتالی کمک گرفته شده و مدل‌ها و تکنیک‌های مختلفی بر روی این داده‌ها مورد آزمایش قرار گرفته‌اند.

۲. پیشینه و کارهای مرتبط

رویه‌های مختلف و متعددی تاکنون برای تحلیل احساسات از روی متن ارائه شده است. این پژوهش‌ها با استفاده از مدل‌های سطحی و عمیق یادگیری به بررسی قطبیت متون پرداخته‌اند. برخی از آنها بر مبنای زبان‌شناسی رایانشی طراحی شده‌اند، اما بیشتر آنها بر پایه یادگیری ماشین بوده و دنباله‌روی پَنگ[9] می‌باشد که تحلیل احساس را از جمله مسائل دسته‌بندی متن در نظر گرفته و از این رو سه روش یادگیری ماشین

[7] User-generated content
[8] Natural Language Processing (NLP)
[9] Pang

با ناظر: بیز ساده[۱۰]، آنتروپی بیشینه[۱۱] و ماشین بردار پشتیبانی[۱۲] را به کار برده است (Pang, et al., 2002). وَن[۱۳] و مَنگ[۱۴] نیز از ویژگی دوتایی‌های واژه‌ها[۱۵] در امر تحلیل احساس استفاده کرده‌اند. آنها همچنین نوع دیگری از ماشین بردار پشتیبانی را پیشنهاد دادند که از نسبت تکرار لگاریتمی استفاده می‌کند و NBSVM نام دارد (Wang & Manning, 2012).

یکی از اولین پژوهش‌های صورت گرفته در زمینه نظرکاوی برای زبان فارسی مربوط به گردآوری مجموعه‌ای با نام PersianClues است. این پژوهش با استفاده از یک روش ابتکاری بدون ناظر که LDASA نام دارد به تحلیل احساس می‌پردازد. در واقع تغییری که در روش پیشنهادی مبتنی بر تخصیص پنهان دیریکله[۱۶] صورت گرفته اضافه کردن مجموعه کلمات حاوی بار معنایی به عنوان بردار ویژگی‌ها در مرحله یادگیری است (Shams, et al., 2012). پژوهش دیگری نیز تحت عنوان ایجاد یک سیستم نظرکاوی با استفاده از الگوریتم‌های با ناظر انجام گرفته است. در گام نخست آن، یک لغت‌نامه احساس برای زبان فارسی به کمک شبکه واژگانی فارسی موجود، فارس نت، گسترش داده شده است. این پژوهش با استفاده از سه الگوریتم یادگیری ماشین، شامل: ماشین بردار پشتیبانی، بیز ساده و رگرسیون منطقی[۱۷] به ارزیابی روش پیشنهادی خود پرداخته است (Basiri, et al., 2014).

رویکرد یادگیری عمیق، به عنوان عرصه‌ای جدید در یادگیری ماشین، توجه بسیاری از محققان را در کاربردهای مختلف به خود جلب کرده است. امروزه یادگیری عمیق در بسیاری از شاخه‌های حوزه پردازش زبان طبیعی همچون تحلیل احساسات، ترجمه ماشین[۱۸] و غیره به کار می‌رود (LeCun, et al., 2015). در این روش‌ها، رایج است کلمات را بصورت بردارهای one-hot بازنمایی کنند که این امر باعث از دست دادن ارتباط ساختاری بین واژه‌ها می‌شود. در بازنمایی‌های پیشرفته‌تر، تشابه بین کلمه‌ها بصورت فاصله‌ای در فضای پیوسته چندبعدی نشان داده می‌شود (Maas, et al., 2011).

یکی دیگر از پژوهش‌های صورت گرفته در زبان فارسی تحت عنوان بهره‌برداری از یادگیری عمیق در تحلیل احساس است. در این پژوهش از مدل یادگیری عمیق شامل شبکه عصبی پیچشی[۱۹] و خودرمزگذار[۲۰] استفاده شده است. و در نهایت مدل یادگیری عمیق معرفی شده خود را با روش‌های کم‌عمق یادگیری ماشین همچون پرسپترون چندلایه[۲۱] مقایسه نموده‌اند (Dashtipour, et al., 2018).

---

۱۰ Naive Bayes (NB)
11 Maximum Entropy
۱۲ Support Vector Machine (SVM)
۱۳ Wan
۱۴ Mang
۱۵ Bigram
۱۶ Latent Dirichlet allocation (LDA)
17 Logistic regression
۱۸ Machine Translation
۱۹ Convolutional Neural Network (CNN)
20 Autoencoder
21 Multilayer perceptron (MLP)

۳. الگوریتم پیشنهادی

مجموعه‌داده[۲۲] مورد استفاده در این مقاله، SentiPers (Hosseini, et al., 2018) نام دارد که نظرات خریداران محصولات از فروشگاه اینترنتی دیجی‌کالا[۲۳] را جمع آوری کرده و برای هر محصول، به ازای هر نظر و هر جمله آن، ویژگی‌هایی نظیر قطبیت، کلمات کلیدی، اهداف و غیره مشخص شده است.

لازم به ذکر است که در این مقاله به دلیل تمرکز بر شاخه تحلیل احساسات در سطح جمله و جلوگیری از پیچیدگی‌های اضافی فقط از ویژگی قطبیت جملات استفاده خواهد شد. قطبیت جملات در این پیکره به‌صورت عددی بین ۲- تا ۲+ نمایش داده شده اند که عدد کوچکتر نشانگر قطبیت کمتر (بار منفی بیشتر) است.

به عنوان نخستین گام، رابطی[۲۴] برای این پیکره پیاده‌سازی و منتشر[۲۵] شده تا تمام جملات و قطبیت مرتبط با آنها را از بین نظرات همه محصولات استخراج و در یک فایل واحد جمع‌آوری نماید.

از مجموع ۷۴۱۵ داده موجود، ۷۵ درصد آن معادل ۵۵۶۱ عدد به عنوان داده یادگیری و ۲۵ درصد باقیمانده معادل ۱۸۵۴ عدد به عنوان داده تست در نظر گرفته شده که انتخاب داده‌ها در طول این تفکیک، به‌صورت تصادفی انجام گرفته است.

۳-۱. تقویت داده

با بررسی تعداد جملات مجموعه‌داده اصلی مشاهده می‌شود که عدم توازنی در میزان داده‌ها به ازای هر دسته (قطبیت) وجود دارد. از آنجایی که این عدم توازن و علاوه بر آن، حجم بسیار پایین داده در یادگیری شبکه عصبی و همچنین در معیارهای ارزیابی بسیار تاثیرگذار است، جهت رفع آن، سه رویکرد مختلف پیشنهاد داده شده و مورد آزمایش و ارزیابی قرار گرفته‌اند. رویکرد نخست بر حفظ توازن و رویکردهای بعدی بر افزایش تعداد داده‌ها تمرکز دارند. تاثیراتی که اعمال هرکدام این رویکردها بر روی مجموعه‌داده و مدل‌های دسته‌بندی داشته‌اند در طول این مقاله مورد بررسی و مقایسه قرار خواهند گرفت. لازم به ذکر است که این روش‌ها فقط بر روی داده‌های یادگیری اعمال شده‌اند و داده‌های تست بدون تغییر در تعداد حفظ خواهند شد تا امکان ارزیابی صحیح تاثیر هرکدام از این روش‌ها با حالت اولیه و با یکدیگر وجود داشته باشد.

۳-۱-۱. تقویت داده: داده‌های اضافه

در این روش از داده‌هایی کمک گرفته شد که به عنوان داده‌های اضافی SentiPers منتشر شده‌اند. ساختار این داده‌ها با پیکره نسخه اصلی تفاوت داشته و از این رو پیاده‌سازی یک استخراج‌کننده دیگر نیز انجام گرفته است. با کاهش بیشینه تعداد داده‌ها در یک دسته و همچنین اضافه کردن داده‌های بیشتر به دسته‌هایی که با مشکل کمبود داده (نسبت به مقدار بیشینه) مواجه بودند، سعی شده تا حد امکان در این تعداد توازن نیز رعایت شود.

---

[22] Dataset

[۲۳] https://www.digikala.com

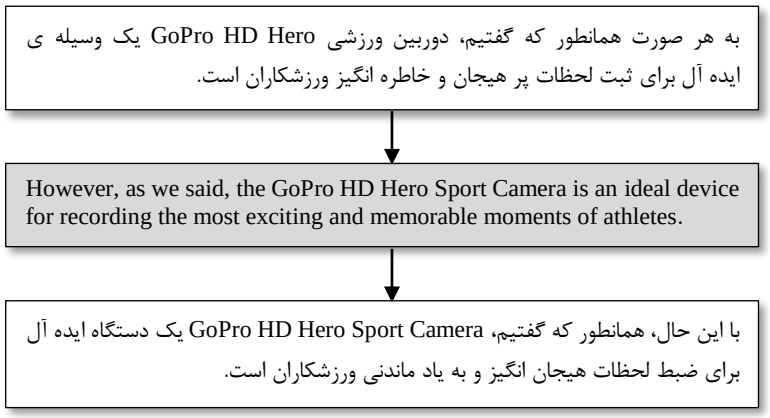
[24] Interface

[25] https://github.com/JoyeBright/Sentiment-Analysis

۲-۱-۳. تقویت داده: ترجمه جملات

در این روش و همچنین روش بعدی که تمرکز آنها بر روی افزایش میزان داده است، از تکنیک‌های نوفه‌دار کردن داده[۲۶] کمک گرفته شده است (Xie, et al., 2017). برای این کار در زمینه‌های تصویر و صدا، تکنیک‌های مختلف و شناخته‌شده‌ای وجود دارد ولی درباره متن با محدودیت‌های دستور زبان، واژگان، مفهوم جمله و غیره مواجه خواهیم شد.

با این حال، از آنجایی که قصد تحلیل احساس جملات را داریم، نقش دستوری کلمات در جمله یا ترتیب آنها بر روی حاصل کار چندان تاثیرگذار نخواهد بود و تمرکز بر روی واژه‌هایی است که عموماً در یک قطبیت خاص به کار رفته‌اند. بنابراین حتی از روی جملات به ظاهر مصنوعی همچون ترجمه‌هایی که توسط مترجم‌گوگل[۲۷] انجام می‌گیرد نیز می‌توان به قطبیت این گونه جمله‌ها پی برد. بنا بر همین فرضیه، هر کدام از جملات موجود در مجموعه‌داده اولیه به کمک مترجم‌گوگل به یک زبان واسط (در اینجا زبان انگلیسی) ترجمه شده و سپس مجدداً از این زبان واسط به زبان فارسی ترجمه شده‌اند (Fadaee, et al., 2017). در فرآیند این ترجمه‌ها، برخی واژه‌ها به واژه‌های مترادف‌شان تبدیل خواهند شد، موقعیت واژه‌ها در جمله تغییر خواهد کرد و بسیاری از تغییرات غیرقابل پیش‌بینی دیگر رخ خواهد داد، اما در نهایت وقتی که کل جمله خوانده می‌شود، مشاهده می‌شود که قطبیت و حس آن حفظ شده است. با اعمال روش بالا، به ازای هر داده موجود، جمله ترجمه شده‌ای نیز به مجموعه‌داده اضافه شده و در نهایت تعداد کل داده‌ها دو برابر می‌شود. در شکل ۱ مثالی از فرآیند تقویت داده به کمک رویکرد ترجمه جملات آمده است.



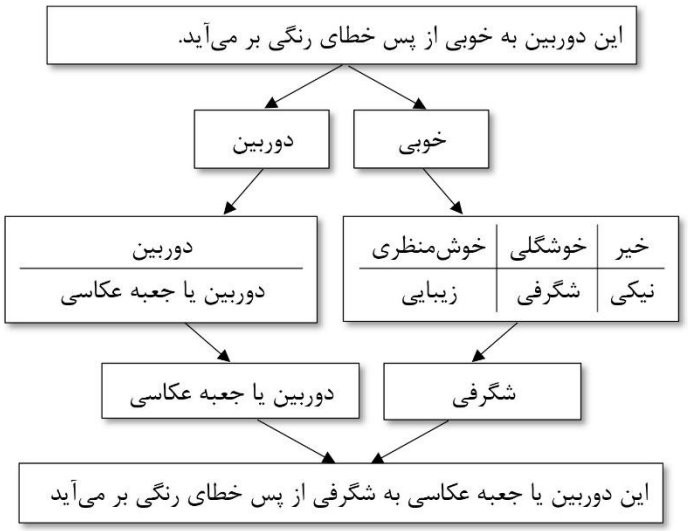
شکل ۱. مثال تقویت داده به کمک ترجمه جملات

۳-۱-۳. تقویت داده: جایگزینی با مترادف

این روش برخلاف روش پیشین که به ترجمه کل جمله می‌پرداخت، تنها چند کلمه موجود در جمله را با مترادف‌شان جایگزین می‌کند. هنگامی که در مترجم‌گوگل واژه‌ای را به زبان دیگر ترجمه می‌کنید، لیستی از واژه‌های مترادف آن و ترجمه مجددش به زبان مبداء نیز به نمایش در می‌آید. در این رویکرد از همین قابلیت

[۲۶] Data noising
[27] Google Translate

مترجم‌گوگل کمک گرفته شده، بدین صورت که ابتدا ۲۰٪ کلمات موجود در هر جمله، بصورت تصادفی جهت جایگزینی انتخاب می‌شوند. لازم به ذکر است که پیش از این انتخاب، برخی پیش‌پردازش‌ها همچون حذف اعداد، تبدیل فاصله به نیم‌فاصله، حذف علائم نگارشی و غیره نیز انجام گرفته است. سپس هرکدام این کلمات توسط این مترجم، به زبان انگلیسی ترجمه شده و لیست کلمات مترادف آن در اختیار قرار می‌گیرد. حال از این لیست، یک کلمه بصورت تصادفی انتخاب شده و به جای کلمه انتخابی اصلی در جمله می‌نشیند. از آنجایی که در دو بخش این روش با انتخاب تصادفی مواجه هستیم با هربار اجرای روند مذکور، ممکن است نتیجه متفاوتی ظاهر گردد. در شکل ۲، نحوه کار این روش قابل مشاهده است. این کار را بر روی تمام جملات تکرار کرده و سرانجام مجموعه‌داده‌ای با دو برابر اندازه اولیه خواهیم داشت.



شکل ۲. مثال تقویت داده به کمک جایگزینی با مترادف

تعداد جملات هر دسته در مجموعه‌داده اصلی و همچنین پس از اعمال هر کدام از روش‌های تقویت داده طبق جدول ۱ به شرح زیر است:

جدول ۱. تعداد جملات مجموعه‌داده

قطبیت	+۲	+۱	۰	-۱	-۲
مجموعه‌داده اصلی	۹۸۸	۱۶۲۳	۲۴۰۹	۵۱۳	۲۸
رویکرد نخست	۱۹۲۸	۲۰۰۰	۲۰۰۰	۹۳۷	۱۵۸
رویکرد دوم/سوم	۱۹۷۶	۳۲۴۶	۴۸۱۸	۱۰۲۶	۵۶

همانطور که گفته شد مجموعه‌داده موجود با توجه به میزان قطبیت جمله در پنج دسته مختلف طبقه‌بندی شده است. بنابراین در حالت اولیه با یک مسئله دسته‌بندی چندگانه[۲۸] مواجه هستیم که پیچیدگی‌های خاص خود را دارا می‌باشد. با این حال در برخی از پژوهش‌ها، شدت قطبیت در نظر گرفته نشده و دسته‌بندی به

[28] Multinomial classification

شکل دودویی[۲۹] انجام می‌گیرد. بدین صورت که برای هر جمله صرفاً یکی از برچسب‌های مثبت یا منفی انتخاب می‌شود. ما نیز در این مقاله به بررسی هر دو روش دسته‌بندی فوق پرداخته و برای هرکدام از آنها، مدل‌های پیشنهادی خود را مورد آزمایش قرار داده‌ایم. جهت نگاشت دسته‌بندی چندگانه به دودویی، جملات کلاس ۱+ و ۲+ در دسته‌ی مثبت و جملات کلاس ۱- و ۲- در دسته‌ی منفی در نظر گرفته می‌شوند. در مورد جملات کلاس ۰ یعنی جملات خنثی می‌توان سه رویکرد را در پیش گرفت. اگر این جملات را به یکی از دسته‌های مثبت یا منفی انتقال دهیم با توجه به عدم توازن در مجموعه‌داده و تعداد بالای جملات مثبت، امکان تحلیل صحیح مدل‌ها و ارزیابی آنها به کمک میزان دقت مدل میسر نخواهد بود. از این رو بهترین رویکرد این است که جملات خنثی در دسته‌بندی دودویی در نظر گرفته نشوند.

۲-۳. الگوریتم‌های پایه

الگوریتم‌های احتمالاتی و یادگیری ماشین که مستقیماً از قوانین ریاضی بهره می‌برند در بسیاری از مسائل حوزه پردازش زبان طبیعی، از خود عملکرد خوبی نشان داده‌اند. این الگوریتم‌ها همچنان و عموماً در مواردی که امکان دسترسی به میزان بالای داده وجود ندارد، مورد استفاده قرار می‌گیرند. از همین رو، ما نیز سه مورد از این الگوریتم‌های شناخته‌شده را بر روی هرکدام از مجموعه داده‌ها و به هر دو شکل دسته‌بندی دودویی و دسته‌بندی چندگانه مورد آزمایش و بررسی قرار خواهیم داد.

۱-۲-۳. بازنمایی کلمات

از آنجایی که تمام مدل‌های مورد استفاده در این بخش مبنای ریاضی دارند، در کار با داده‌های متنی نیاز به روشی است تا بتوان این رشته‌های متنی را به اعداد تبدیل کرد. در مدل‌های پایه هر کدام از کلمات موجود در متن را با مقدار tf-idf[۳۰] آن تعبیه[۳۱] خواهیم کرد (Ramos, 2003). مقادیری که از این تکنیک بدست می‌آیند به خوبی نشانگر اهمیت هر کلمه نسبت به یک سند در کل پیکره می‌باشند و به کمک آن می‌توان کلماتی که بصورت کلی بیشتر تکرار شده‌اند را پیدا کرد.

در حین اجرای رویه تعبیه‌سازی، عمل پیش‌پردازش متن نیز انجام خواهد شد. بدین صورت که ابتدا کلمات موجود در هر جمله به کمک ابزار هضم[۳۲] که منحصراً برای زبان فارسی پیاده‌سازی شده، تفکیک می‌شوند. سپس از تکنیک N-تایی (Sugathadasa, et al., 2018) با مقدار N=2 و در سطح کلمه نیز استفاده می‌شود. با این کار جفت کلماتی که معمولاً کنار همدیگر ظاهر می‌شوند نیز در الگوریتم در نظر گرفته خواهند شد. بنابراین در پایان کار مقدار tf-idf هم برای تک تک کلمات و هم برای هر جفت کلمه بدست خواهد آمد.

علاوه بر این، عبارت‌هایی (تک کلمه‌ها و جفت کلمه‌ها) که در بیش از یک جمله به کار برده نشده‌اند در طول عمل شمارش و سپس ساخت بردارهای tf-idf در نظر گرفته نمی‌شوند.

---

[29] Binary classification
[30] Term frequency–inverse document frequency
[31] Embedding
[32] https://github.com/sobhe/hazm

۲-۲-۳. مدل‌های پایه

در رویکرد یادگیری ماشین، تحلیل احساس با استفاده از قواعد زبانی یا نحوی و یا هر دو به عنوان یک مسئله طبقه‌بندی متن عادی تلقی شده و الگوریتم‌های معروف یادگیری ماشین بر روی آن اعمال می‌شود. ما نیز در این مقاله به استفاده از سه الگوریتم معروف بیز ساده (Thorsten, 1997)، گرادیان کاهشی تصادفی[۳۳] (Prasetijo, et al., 2017) و ماشین بردار پشتیبانی (Li & Li, 2013) که تاکنون در زمینه طبقه‌بندی متن به خوبی عمل کرده‌اند به عنوان مدل‌های پایه استفاده خواهیم کرد.

این سه الگوریتم بر روی هرکدام از مجموعه‌داده‌ها، یکبار به‌صورت دسته‌بندی چندگانه و یکبار دسته‌بندی دودویی اعمال شده و نتایج حاصل در جداول ۲ و ۳ آمده است:

جدول ۲. دقت مدل‌های پایه بر اساس دسته‌بندی چندگانه (برحسب درصد)

مدل / مجموعه‌داده	SVM	SGD	NB
اولیه	۶۴/۵۶	۵۹/۳۸	۴۸/۷۵
تقویت‌یافته – داده‌های اضافه	۶۷/۶۹	۶۶/۱۸	۵۸/۲۵
تقویت‌یافته – ترجمه جملات	۶۵/۴۲	۵۷/۹۲	۵۲/۲۶
تقویت‌یافته – جایگزینی با مترادف	۶۵/۵۸	۵۷/۸۲	۵۰/۱۰

جدول ۳. دقت مدل‌های پایه بر اساس دسته‌بندی دودویی (برحسب درصد)

مدل / مجموعه‌داده	SVM	SGD	NB
اولیه	۸۶/۴۰	۸۲/۲۶	۸۲/۳۵
تقویت‌یافته – داده‌های اضافه	۹۱/۸۰	۸۲/۲۶	۸۲/۸۰
تقویت‌یافته – ترجمه جملات	۸۷/۶۶	۸۲/۳۵	۸۲/۴۴
تقویت‌یافته – جایگزینی با مترادف	۸۷/۲۱	۸۲/۳۵	۸۲/۳۵

۳-۳. یادگیری عمیق

یادگیری عمیق برای نخستین بار توسط هیلتُن[۳۴] در سال ۲۰۰۶ و به عنوان بخشی از فرآیند یادگیری ماشین که به شبکه عصبی عمیق مرتبط می‌شد مطرح گردید (Day, 2016). شبکه‌های عصبی قادر هستند مجموعه‌داده را به روش با ناظر و بدون ناظر آموزش و طبقه‌بندی کنند (Vateekul & Koomsubha, 2016). یادگیری عمیق شامل شبکه‌های زیادی از جمله: شبکه‌های عصبی پیچشی، شبکه‌های عصبی بازگشتی[۳۵] و

[۳۳] Stochastic Gradient Descend (SGD)

[۳۴] G.E. Hilton

[35] Recursive Neural Networks (RNN)

شبکه‌های باور عمیق[۳۶] و بسیاری مدل‌های دیگر است و کاربرد فراوانی در نمایش برداری، تخمین نمایش کلمه، مدل سازی جملات و نمایش ویژگی‌ها دارد.

یادگیری عمیق اخیراً برای حل بسیاری از مسائل پردازش زبان طبیعی مورد استفاده قرار گرفته است (Collobert, et al., 2011). اصلی‌ترین مزیت این رویکرد، عدم نیاز به تنظیم دستی و استخراج ویژگی‌ها بر اساس دانش تخصصی و دسترسی به منابع زبانی است (Rojas-Barahona & Maria, 2016).

در این مقاله با توجه به بهره بردن شبکه‌های عصبی حافظه طولانی کوتاه مدت[۳۷] از حافظه داخلی و امکان نگهداری داده‌های پیشین و همچنین استخراج ویژگی‌های محلی حول پنجره‌ای در شبکه‌های عصبی پیچشی (Collobert, et al., 2011)، از ساختارهایی بر پایه این مدل‌ها و مناسب طبقه‌بندی متن استفاده خواهیم کرد.

۳-۳-۱. پیش‌پردازش

در این مرحله به کمک ابزار هَضم، نسبت به پیش‌پردازش متون اقدام کرده‌ایم. روند پیش‌پردازش بدین صورت است که ابتدا بر روی متن موجود، نرمال‌سازی انجام می‌گیرد که طی این مرحله در برخی موارد لازم، فاصله‌ها به نیم‌فاصله تبدیل می‌شوند. سپس متن نرمال شده، نشانه‌گذاری شده و به لیستی از کلمات تبدیل می‌شود. حال از این لیست هر جزء که برابر حروف زائد یا عدد بوده یا طول کمتر از ۱ داشته باشند حذف خواهند شد. سپس کلمات باقیمانده توسط ریشه‌یاب این ابزار، به ریشه خود تبدیل شده و در آخر کلمات این لیست دوباره گرد هم آمده و به شکل یک جمله در می‌آیند. لازم به ذکر است استفاده از ریشه‌یاب به دلیل کاهش پراکندگی کلمات هم‌ریشه و بالابردن دقت حاصل از مدل‌ها می‌باشد. این روند پیش‌پردازش بر روی تمام داده‌های یادگیری اعمال خواهد شد.

۳-۳-۲. بازنمایی کلمات

همانطور که در بخش (۳,۲,۱) گفته شد نیاز است متون مجموعه‌داده به‌صورت ساختاری عددی تعبیه شوند. در این روش از امکاناتی که یادگیری عمیق و شبکه‌های عصبی در اختیارمان می‌گذارد استفاده کرده و برخلاف تعبیه تک‌بعدی کلمات در مدل‌های پایه، در اینجا از فضای چندبعدی کمک خواهیم گرفت. بدین صورت که هر متن به شکل مجموعه‌ای از کلمات درآمده و سپس هر کلمه را به برداری در فضای چندبعدی نگاشت خواهیم داد. مقدار هر بعد با یک ویژگی خاص مطابقت دارد و می‌تواند حتی تفسیری معنایی[۳۸] یا دستوری[۳۹] داشته باشد که آن را ویژگی کلمه[۴۰] می‌نامیم (Turian, et al., 2010). با استفاده از این روش، این امکان به وجود می‌آید کلماتی که از لحاظ مفهوم یا ظاهر شدن در جمله به یکدیگر نزدیک‌اند، در فضای چند بعدی نیز با بردارهایی نزدیک به یکدیگر تعبیه شوند.

عمل یادگیری و تعبیه این بردارها به دو طریق ممکن خواهد بود که به بررسی و مقایسه هرکدام آنها خواهیم پرداخت. روش نخست بر پایه مجموعه‌داده موجود بوده و طی یادگیری شبکه عصبی انجام می‌گیرد.

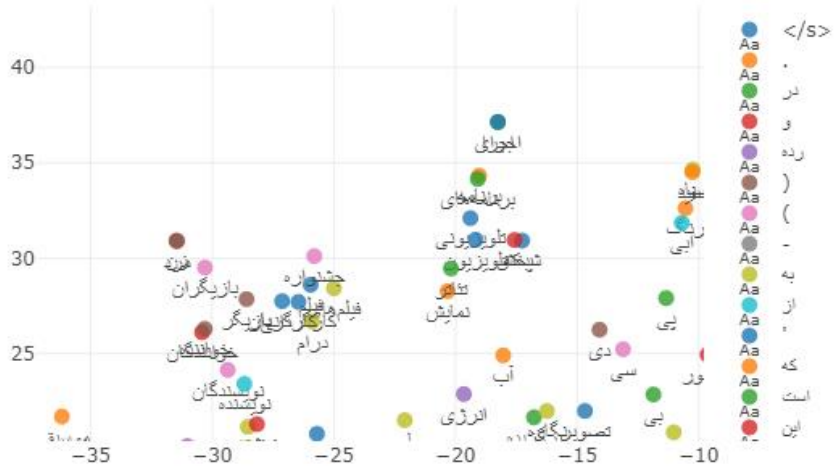
---

[36] Deep Belief Networks (DBN)
[37] Long short-term memory (LSTM)
[38] Semantic
[39] Grammatical
[40] Word Feature

در این روش شبکه عصبی علاوه بر وظایف اصلی، در لایه نخست خود به جستجوی الگوها بر روی مجموعه‌داده پرداخته و طبق این الگوها کلمات را به شکل بردارها تعبیه خواهد کرد. طبعاً این روش به میزان داده موجود و رخداد کلمات در داده‌های مختلف بستگی خواهد داشت. جهت استفاده از این روش از لایه تعبیه‌گر Keras استفاده خواهیم کرد و آن را به همین نام خواهیم خواند.

اما در روش دوم، از بردارهای از پیش آموزش‌دیده[41] استفاده خواهد شد و دیگر شبکه عصبی بر روی لایه ورودی خود عمل یادگیری را انجام نمی‌دهد. شرکت‌ها و دانشگاه‌های مختلفی تاکنون نسبت به آموزش شبکه‌های عصبی بر روی مجموعه‌داده‌های بزرگ و سرانجام ارائه این بردارهای تعبیه اقدام کرده‌اند که برای زبان فارسی، تاکنون بهترین گزینه موجود، کتابخانه FastText[42] از شرکت فیسبوک می‌باشد. این کتابخانه جهت آموزش خود از داده‌های وب‌سایت ویکی‌پدیا استفاده کرده و هر کلمه یا مجموعه‌کلمات را به برداری در فضای ۳۰۰ بعدی نگاشت می‌دهد.

همانطور که در شکل ۳ دیده می‌شود، کلماتی که به یک مفهوم مشترک مربوط هستند از موقعیت‌های نزدیک‌تری به یکدیگر برخوردارند. به عنوان مثال، در این شکل، کلماتی مثل "جشنواره"، "فیلم"، "بازیگر"، "کارگردان" در کنار هم دیده می‌شوند و کل این مجموعه نیز در همسایگی مجموعه دیگری که شاملی همچون کلمات "نمایش"، "تلویزیون"، "تئاتر" است، قرار دارد.



شکل ۳. بخشی از نقشه تعبیه کلمات به کمک FastText

۳-۳-۳. مدل‌ها

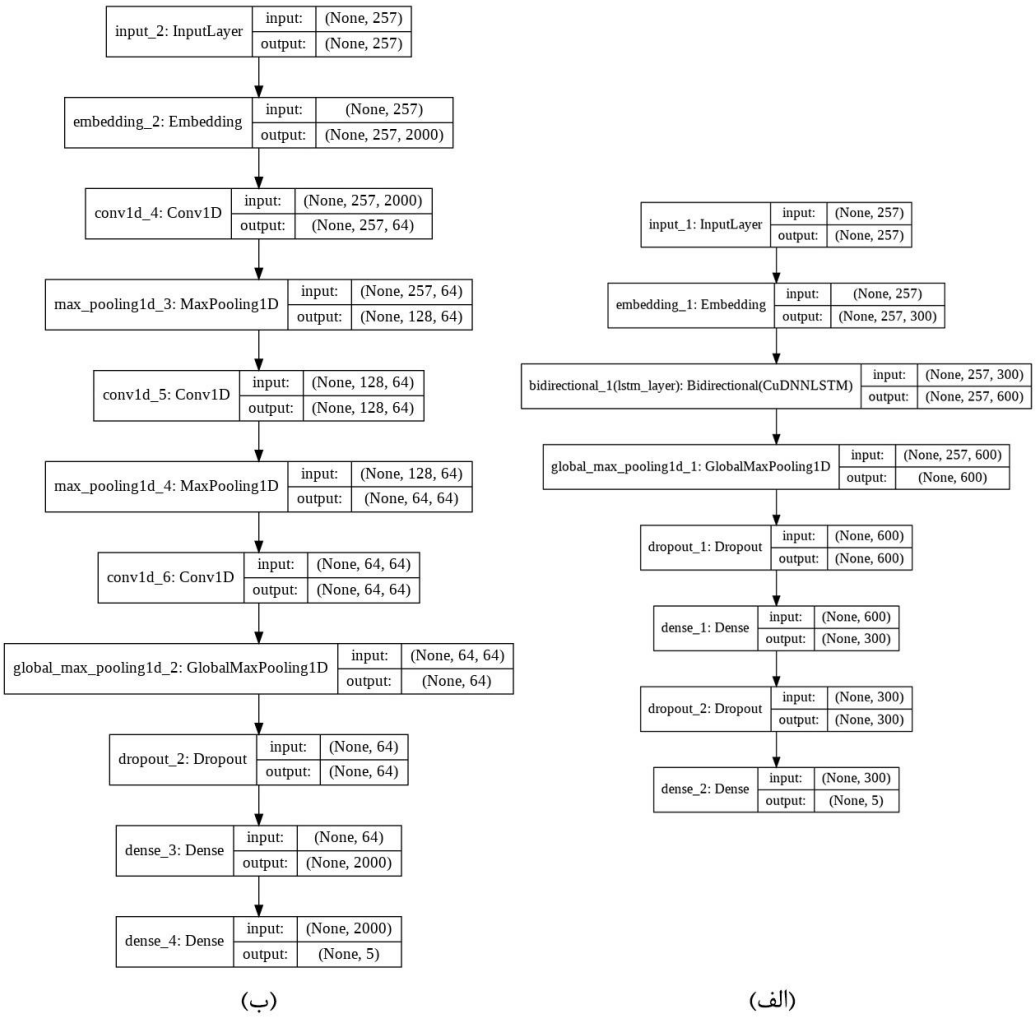
از آنجا که هر متن به‌صورت مجموعه‌ای از کلمات تعبیه‌شده در آمده، در لایه اول مدل‌های شبکه عصبی خود، تعداد نرون‌ها به اندازه بیشترین طول جملات بر حسب کلمه است. در این مجموعه‌داده، طولانی‌ترین متن موجود شامل ۲۵۷ کلمه بوده و بنابراین در لایه اول ۲۵۷ نرون خواهیم داشت. لایه بعدی مدل‌ها، لایه تعبیه کلمات است که هر کلمه را به شکل برداری در فضای چند بعدی تعبیه می‌کند. در مدل‌هایی که از FastText

[41] Pre-trained Word Embedding
[42] https://fasttext.cc/docs/en/crawl-vectors

استفاده شده این ابعاد از قبل تعیین شده و برابر ۳۰۰ بعد می‌باشد. در تعبیه کلمات به کمک Keras، از آنجا که یادگیری تعبیه در خود شبکه عصبی صورت می‌گیرد این ابعاد ۲۰۰۰ در نظر گرفته شده است.

در این مدل‌ها از تکنیک حذف تصادفی[۴۳] نیز استفاده شده که هر مرتبه به‌صورت تصادفی درصدی از وزن‌های شبکه عصبی را در طول فرآیند یادگیری حذف کرده و در نتیجه از بیش‌برازش[۴۴] مدل‌ها می‌کاهد (Srivastava, et al., 2014).

در این قسمت از دو ساختار مختلف برای لایه‌های شبکه عصبی استفاده شده است. نخستین ساختار حافظه طولانی کوتاه مدت دوطرفه[۴۵] است که بر پایه شبکه‌های عصبی بازگشتی طراحی شده و دوطرفه بودن آن امکان دریافت اطلاعات توسط گذشته و آینده را به لایه خروجی آن اضافه می‌کند. در شکل ۴-الف، لایه‌های تعبیه‌شده برای این ساختار قابل مشاهده است.



شکل ۴. ساختار لایه‌ها در الف) مدل BI- LSTM و تعبیه کلمات به کمک FastText
ب) مدل CNN و تعبیه کلمات به کمک Keras

[۴۳] Dropout
[۴۴] Overfitting
[۴۵] Bidirectional Long Short-Term Memory (BI-LSTM)

ساختار بعدی مورد استفاده در این مقاله شبکه عصبی پیچشی نام دارد که یکی از موفق‌ترین ساختارهای شبکه‌های عصبی برای حوزه پردازش تصویر است اما کیم[۴۶] نشان داد که این مدل برای داده‌های متنی خصوصاً در مسائل طبقه‌بندی متن نیز به خوبی عمل می‌کند (Kim, 2014). لایه‌های استفاده شده در این ساختار در شکل ۴-ب قابل مشاهده است.

نتیجه آموزش این شبکه‌های عصبی بر روی هرکدام از مجموعه‌داده‌ها و پیش‌بینی توسط آنها در جدول ۴ برای دسته‌بندی چندگانه و در جدول ۵ برای دسته‌بندی دودویی آورده شده است.

جدول ۴. دقت مدل‌های یادگیری عمیق بر اساس دسته‌بندی چندگانه (برحسب درصد)

مدل / مجموعه‌داده	CNN FastText	CNN Keras	LSTM FastText	LSTM Keras
اولیه	۵۳/۸۲	۵۹/۹۲	۶۳/۳۷	۶۰/۹۴
تقویت‌یافته – داده‌های اضافه	۶۴/۵۶	۶۶/۳۴	۶۷/۷۹	۶۷/۸۵
تقویت‌یافته – ترجمه جملات	۵۹/۶۵	۶۴/۹۴	۶۷/۴۲	۶۵/۲۶
تقویت‌یافته – جایگزینی با مترادف	۵۷/۰۱	۶۵/۱۰	۶۳/۴۸	۶۵/۳۱

جدول ۵. دقت مدل‌های یادگیری عمیق بر اساس دسته‌بندی دودویی (برحسب درصد)

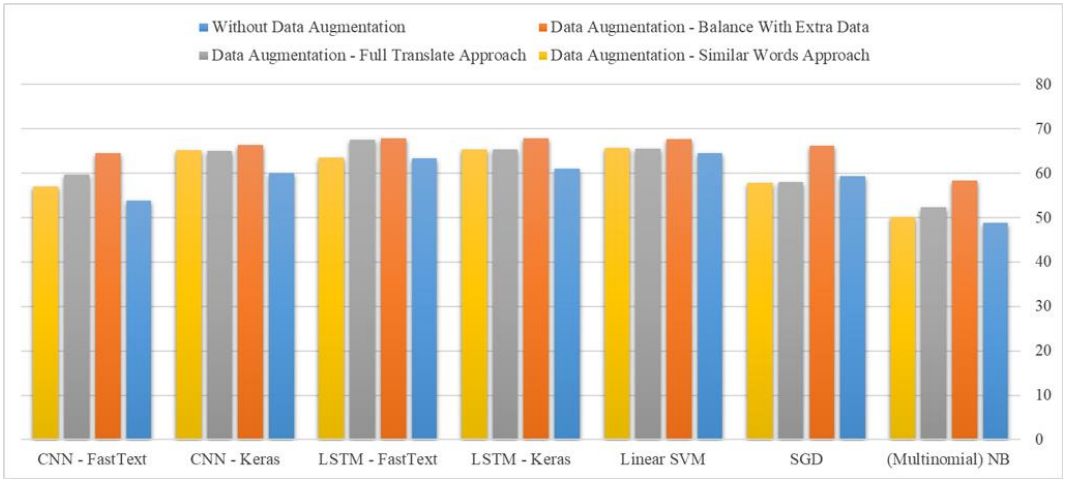
مدل / مجموعه‌داده	CNN FastText	CNN Keras	LSTM FastText	LSTM Keras
اولیه	۸۱/۰۰	۸۷/۰۳	۸۵/۴۱	۸۵/۹۵
تقویت‌یافته – داده‌های اضافه	۸۲/۵۳	۹۱/۸۰	۹۰/۲۷	۹۰/۲۷
تقویت‌یافته – ترجمه جملات	۷۹/۹۲	۸۷/۴۸	۸۷/۷۵	۸۵/۳۲
تقویت‌یافته – جایگزینی با مترادف	۸۱/۲۷	۸۶/۸۵	۸۶/۲۲	۸۶/۳۱

۴. نتایج

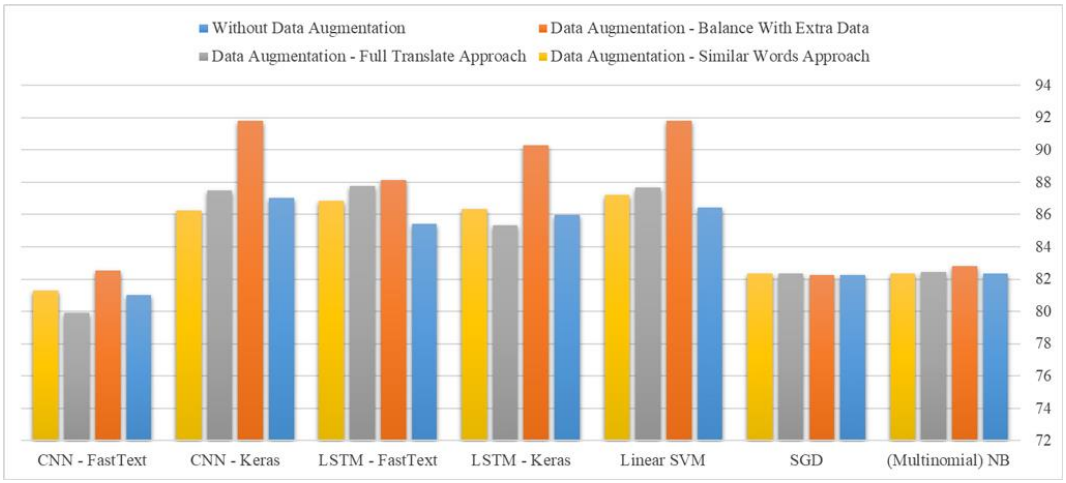
نتیجه اعمال تمام مدل‌ها اعم از الگوریتم‌های یادگیری ماشین و مدل‌های یادگیری عمیق به‌صورت دسته‌بندی چندگانه و بر روی هرکدام از مجموعه‌داده‌ها در نمودار ۱ قابل ملاحظه است.

همچنین درصد دقت بدست آمده به‌صورت دسته‌بندی دودویی برای تمام مدل‌ها و بر روی هر کدام از مجموعه‌داده‌ها در نمودار ۲ آمده است.

[۴۶] Kim



نمودار ۱. دقت مدل‌ها با توجه به هر مجموعه‌داده – دسته‌بندی چندگانه



نمودار ۲. دقت مدل‌ها با توجه به هر مجموعه‌داده – دسته‌بندی دودویی

۵. تحلیل نتایج

از هر دو نمودار ۱ و ۲ می‌توان چنین استنباط کرد که بیش از همه رویکرد اول تقویت داده در بالابردن دقت مدل‌ها موثر بوده است. علت این تاثیر را می‌توان هم توازن صورت گرفته در تعداد داده‌های موجود در هر دسته و هم واقعی و کاملاً جدید بودن این داده‌های اضافه برای مدل‌ها دانست.

پس از آن مشاهده می‌شود که تمام روش‌های پیشنهادی این مقاله در امر تقویت داده توانسته‌اند به میزان قابل‌توجهی میزان دقت دسته‌بندی را نسبت به حالت عادی افزایش دهند.

با بررسی نتایج مدل‌های پایه در این مقاله، مشاهده می‌شود مدل SVM در تمام حالت‌های مورد آزمایش، از عملکرد بهتری برخوردار بوده و دقت آن نسبت به سایر الگوریتم‌های مورد استفاده به میزان قابل توجهی بیشتر است. و همچنین از نتایج مدل‌های یادگیری عمیق این چنین برداشت می‌شود که تعبیه کلمات به کمک FastText همواره در مدل BI-LSTM بهتر از مدل CNN عمل کرده است. همچنین در دسته‌بندی دودویی مدل CNN با تعبیه‌گر Keras و در دسته‌بندی چندگانه مدل BI-LSTM با تعبیه به کمک FastText بیشترین دقت را نسبت به سایرین کسب کرده‌اند.

۶. جمع‌بندی و کارهای آتی

عموماً مدل‌های یادگیری عمیق نتیجه بهتر یا قابل‌قبولی را کسب کرده اند و در صورت وجود مجموعه‌داده با اندازه و توازن مناسب، که هسته اصلی اینگونه مدل‌ها به شمار می‌روند امکان بهبود این میزان دقت تا حد بالایی وجود دارد.

روش‌های ارائه شده برای تقویت داده از تاثیر چشمگیری در نتایج برخوردار بوده‌اند و انتظار می‌رود اعمال این روش‌ها بر روی سایر مجموعه‌داده‌های متنیِ محدود و افزایش تعداد داده‌های آنها، افزایش عملکرد مدل‌ها را حداقل برای کاربرد دسته‌بندی متون، در پی داشته باشد. آزمایش این مدل‌ها بر روی مجموعه‌داده‌های بزرگتر زبان فارسی که تاکنون در دسترس نبوده‌اند، تغییر در پارامترهای شبکه‌های عصبی با توجه به این مجموعه‌داده‌ها، تغییر در ساختار لایه‌ها، استفاده از بردارهای نگاشت قوی‌تر از جمله کارهایی است که می‌تواند در آینده مورد تحقیق و بررسی قرار گیرد.

منابع

Basiri, M. E., Nilchi, A. R. N. & Ghassem-aghaee, N., 2014. A Framework for Sentiment Analysis in Persian.

Collobert, R. et al., 2011. Natural Language Processing (Almost) from Scratch. *The Journal of Machine Learning Research*, Volume 12, pp. 2493-2537.

Dashtipour, K. et al., 2018. *Exploiting Deep Learning for Persian Sentiment Analysis.* s.l., s.n.

Day, M., 2016. Deep Learning for Financial Sentiment Analysis on Finance News Providers.

Fadaee, M., Bisazza, A. & Monz, C., 2017. Data Augmentation for Low-Resource Neural Machine Translation. *arXiv*.

Hosseini, P. et al., 2018. SentiPers: A Sentiment Analysis Corpus for Persian. *arXiv*.

Kim, Y., 2014. *Convolutional Neural Networks for Sentence Classification.* Doha, Qatar, s.n.

LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep learning. *Nature*, Volume 521, pp. 436-444.

Liu, B., 2012. Sentiment Analysis and Opinion Mining. *Synthesis lectures on human language technologies*, pp. 1-167.

Liu, B., 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* s.l.:Cambridge University Press.

Li, Y.-M. & Li, T.-Y., 2013. Deriving market intelligence from microblogs. *Decision Support Systems*, 55(1), pp. 206-217.

Maas, A. L. et al., 2011. Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142-150.

Pang, B., Lee, L. & Vaithyanathan, S., 2002. Thumbs up?: sentiment classification using machine learning techniques. *Proceeding,* pp. 79-86.

Prasetijo, A. B. et al., 2017. *Hoax detection system on Indonesian news sites based on text classification using SVM and SGD.* s.l., s.n., pp. 45-49.

Ramos, J., 2003. Using TF-IDF to Determine Word Relevance in Document Queries. *Arxiv.*

Rojas-Barahona & Maria, L., 2016. Deep learning for sentiment analysis, Language and Linguistics Compass. *Language and Linguistics Compass.*

Shams, M., Shakery, A. & Faili, H., 2012. *A non-parametric LDA-based induction method for sentiment analysis.* Shiraz, Iran, s.n.

Srivastava, N. et al., 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research,* Volume 15, pp. 1929-1958.

Sugathadasa, K., Ayesha, B., de Silva, N. & Perera, A., 2018. Legal Document Retrieval using Document Vector Embeddings and Deep Learning. *ArXiv.*

Thorsten, J., 1997. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., pp. 143--151.

Turian, J., Ratinov, L. & Bengi, Y., 2010. *Word representations: a simple and general method for semi-supervised learning.* Stroudsburg, s.n.

Vateekul, P. & Koomsubha, T., 2016. *A Study of Sentiment Analysis Using Deep Learning Techniques on Thai Twitter Data.* Khon Kaen, s.n.

Wang, S. & Manning, C. D., 2012. Baselines and Bigrams : Simple, Good Sentiment and Topic Classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers,* Volume 2, pp. 90-94.

Xie, Z. et al., 2017. Data Noising as Smoothing in Neural Network Language Models. *ICLR.*