ششمین کنفرانس ملی
محاسبات توزیعی و پردازش داده های بزرگ
۱۳ اسفند ۱۳۹۹
6th National Conference on Distributed Computing and Big Data Processing

تحلیل نرخ دلار در شبکه اجتماعی توییتر با استفاده از الگوریتم های یادگیری ماشین

صبا فرهادی [۱]، محبوبه شمسی [۲]، عبدالرضا رسولی کناری [۳]

[۱]دانشجوی مهندسی کامپیوتر، دانشگاه صنعتی قم،

[۲] استادیار مهندسی کامپیوتر، دانشگاه صنعتی قم، shamsi@Qut.ac.ir

[۳] استادیار مهندسی کامپیوتر، دانشگاه صنعتی قم، Rasouli@qut.ac.ir

چکیده

در سال‌های اخیر، تحقیقات تحلیل احساسات در توییتر، توییت‌ها را برای استخراج احساسات کاربر تحلیل می‌کنند که گسترش زیادی داشته است. بسیاری از محققان برای چنین تحلیل‌هایی تصمیم می‌گیرند از الگوریتم‌های یادگیری ماشین و یادگیری عمیق استفاده کنند. در این پژوهش با استفاده از الگوریتم‌های یادگیری ماشین و یادگیری عمیق، یک تحلیل احساسات جزئی از داده‌های توییتر فارسی مربوط به «نرخ دلار» در بازه زمانی سه ماهه تابستان ۱۳۹۷ انجام دادیم. این پژوهش شامل پیش‌پردازش توییت توسط یک رویکرد استخراج ویژگی است. سپس با طبقه‌بندی احساسی شامل ناامیدی، خوشحالی، ترس، ناراحتی و امید، ویژگی‌ها طبقه‌بندی می‌شوند. در چارچوب پیشنهادی، الگوریتم‌های رگرسیون لجستیک، ماشین بردار پشتیبان، درخت تصمیم و جنگل تصادفی و روش جدید ترکیبی رگرسیون لجستیک و ماشین بردار پشتیبان برای ارزیابی تحلیل احساسات مورد استفاده قرار می‌گیرد. همچنین از دو الگوریتم یادگیری عمیق شبکه عصبی پیچشی و شبکه عصبی بازگشتی استفاده شده است. مشاهدات تجربی نشان می‌دهد که احساسات مردم در مورد تغییر نرخ دلار در بازه زمانی سه ماهه تابستان بیشتر به سمت ناامیدی گرایش داشته است. همچنین روش ترکیبی ارائه شده برای ماه اول تابستان با مقدار ۹۹ درصد و دو ماه دیگر با مقدار ۹۷ درصد بیشترین دقت را نسبت به دیگر الگوریتم‌ها داشته است.

کلمات کلیدی:

توییتر، یادگیری ماشین، یادگیری عمیق، احساسات، دلار، دقت

۱- مقدمه

تحلیل احساسات[۱] گاهی اوقات از آن به نظرکاوی[۲] نیز یاد می‌کنند و یک حوزه تحقیقاتی است که هدف آن تحلیل احساسات یا نظرات افراد راجع به موجودیت‌هایی مثل عناوین، رویدادها، افراد خاص، مسائل، خدمات، محصولات، سازمان‌ها و موارد مرتبط با آن‌ها است. تحلیل احساسات شاخه‌ای از یادگیری ماشین، داده‌کاوی[۳]، پردازش زبان طبیعی و زبان‌شناسی محاسباتی است، همچنین مفاهیمی را از علم جامعه‌شناسی

و روان‌شناسی قرض می‌گیرد. از چند سال گذشته، رونق شبکه‌های اجتماعی، توسعه تحلیل احساسات را سرعت بخشیده است. [۱]

تشخیص ساختار متن به سه شکل مختلف سند، جمله و کلمه امکان‌پذیر است. در سطح سند، اسناد به نظرات مثبت، منفی و خنثی طبقه‌بندی می‌شوند. در سطح جمله تعیین می‌کند که آیا جمله نظر مثبت، منفی و یا خنثی را بیان می‌کند و در سطح ویژگی، برای تعیین نظر مثبت، منفی و یا خنثی به جزئی‌ترین شکل ممکن (کلمه) انجام می‌گیرد و نظرات را در مورد ویژگی‌های خاص موضوع بررسی می‌کنیم [۲]. در این مطالعه، طبق داده‌های جمع‌آوری شده توییتر[۴] با هشتگ دلار، تحلیل احساسات در سطح جمله مورد توجه قرار گرفته است.

در گذشته برای به دست آوردن نظرات کاربران از فرم‌های نظرسنجی استفاده می‌شد، ولی امروزه با گسترش وب و توسعه اینترنت این کار به راحتی از طریق پست گذاشتن در شبکه‌های اجتماعی و یا ارائه دیدگاه ذیل هر پست قابل انجام است.رشد اطلاعات به صورت آنلاین در شبکه‌های اجتماعی، تحلیل احساسات را ضروری‌تر کرده است. از لحاظ اقتصادی، تحلیل احساسات می‌تواند توصیه‌ها و پیشنهادات آنلاین برای مشتریان و فروشنده‌ها داشته باشد. از سوی دیگر، این ترجیحات کاربری که داده‌ها نمایان می‌کنند می‌تواند به بسترهای فروش آنلاین کمک کند تا محصولات و خدمات‌شان را تحلیل کنند. از سوی دیگر، به علت ذات مجازی خرید آنلاین، بررسی جزئیات دقیق یک کالا و اینکه آیا مصرف‌کننده مایل است نظرات یا دیدگاه‌های سایر مصرف‌کننده‌ها را بداند آسان نیست.

از لحاظ سیاسی، تقاضاهای گسترده برای اطلاعات سیاسی می‌تواند به یک عامل مهم دیگر در نظر گرفته شود. کاربردهای تجاری، تنها انگیزه مردم برای بررسی و بیان دیدگاه‌ها به صورت آنلاین نیست. به عنوان مثال در تحلیل بحث‌ها در توییتر قبل از انتخابات پارلمان اروپا، محققان بیش از ۱.۲ میلیون توییت به سه زبان (انگلیسی، فرانسوی و آلمانی) در طول دوره دو هفته‌ای جمع‌آوری کردند و میزان مثبت و منفی بودن نظرات مردم نسبت به انتخابات پارلمان را تحلیل کردند. [۱] تحلیل احساسات توییتر در مقایسه با سایر منابع بخاطر وجود لغات عامیانه، اشتباهات تایپی و لغات حاوی طعنه و کنایه، دشوارتر است. ماکزیمم تعداد کاراکتر مجاز در توییتر ۱۴۰ است. روش مبتنی بر پایگاه دانش و یادگیری ماشین، دو استراتژی به‌کار رفته در تحلیل احساسات متون هستند.

در این پژوهش، پست‌های توییتر در رابطه با نرخ دلار را از اطلاعات بی‌معنا می‌زداییم و سپس بر اساس لغت‌نامه احساسی حاوی پنج احساس ناامیدی، خوشحالی، ترس، ناراحتی و امید طبقه‌بندی می‌کنیم. سپس متون را به ماتریس تبدیل کرده و توسط الگوریتم‌های مختلف یادگیری ماشین و یادگیری عمیق و یک روش جدید ترکیبی برای طبقه‌بندی این توییت‌ها به کار برده شده است. سپس برای ارزیابی روش پیشنهادی پارامترهایی چون صحت[۵]، بازیابی[۶]، F-Measure و دقت[۷] را مورد توجه قرار داده‌ایم [۳].

۱-۱- کارهای گذشته

تحقیقات متعددی بر روی تحلیل داده‌های شبکه‌های اجتماعی تمرکز کرده‌اند، خصوصا داده‌هایی که مختص به یک رویداد خاص هستند. این استفاده گسترده از شبکه‌های اجتماعی توجه زیادی را از سوی محققان به خود جلب کرده و بسیاری از تحقیقات برای دریافت اطلاعات مهم بر روی این رویدادها انجام شده است. در سال‌های اخیر تحقیقات گسترده‌ای در زمینه تحلیل احساسات در توییتر انجام شده است.

Jain و Dandannavar چندین مرحله برای تحلیل احساسات بر روی داده‌های توییتر با استفاده از الگوریتم‌های یادگیری ماشین را بررسی کردند. رویکرد آن‌ها در ابتدا داده‌ها را جمع‌آوری می‌کرد و بعد با استفاده از تکنیک‌های NLP پیش‌پردازش می‌کند. سپس، برای استخراج ویژگی‌های مربوط به احساس، عمل استخراج ویژگی بر روی آن‌ها انجام شد. در نهایت، یک مدل با استفاده از دسته‌بندهایی مانند بیز ساده[۸]، ماشین بردار پشتیبان و درخت تصمیم آموزش داده شد. فریمورک ارائه شده تحلیل احساسات را با استفاده از بیز ساده چندجمله‌ای[9] و درخت تصمیم اجرا کرد. نتایج نشان داد که درخت تصمیم موثرتر عمل می‌کند و صحت، دقت، بازیابی و F1-score بهتری را ارائه می‌دهد[۴].

تحقیقات گسترده‌ای نیز توسط Go و همکارانش انجام شده است که یک راه‌حلی با استفاده از نظارت از راه دور[۱۰] برای تحلیل احساسات بر اساس توییت‌ها ارائه داده‌اند. در این روش، آن‌ها از داده‌های آموزشی شامل توییت‌های حاوی شکلک استفاده کردند که به عنوان برچسب‌های پرت[۱۱] عمل می‌کرد. آن‌ها مدل‌هایی بر اساس دسته‌بند بیز ساده، حداکثر آنتروپی، و ماشین بردار پشتیبان ایجاد کردند. ویژگی‌های آن‌ها از یونیگرام[۱۲]، بایگرام[۱۳] و POS[14] تشکیل می‌شد. آن‌ها نتیجه گرفتند که ماشین بردار پشتیبان نسبت به سایر مدل‌ها بهتر عمل می‌کرد و این که یونیگرام‌ها ویژگی موثرتری نسبت به بقیه بودند[۵].

علاوه‌براین، چندین پژوهش در زمینه SemEval، وظیفه دسته‌بندی توییت‌ها با چند صد شرکت‌کننده را مورد بررسی قرار دادند[۶]. در طول

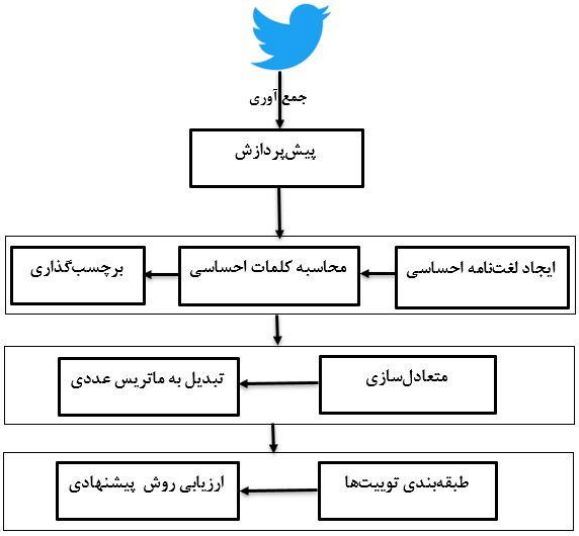
دهه گذشته توجه گسترده‌ای به تحلیل احساسات بر اساس داده‌های توییتر و همچنین رگرسیون ترتیبی وجود داشته است. مسئله رگرسیون، یکی از اصلی‌ترین حوزه‌های پژوهش در زمینه یادگیری ماشین و داده‌کاوی است.

ساختار مقاله به شرح زیر می‌باشد: بخش ۲، پیاده‌سازی روش پیشنهادی تحلیل احساسات توضیح داده شده است. بخش ۳، ارزیابی روش پیشنهادی انجام شده است. بخش ۴، مقاله را به پایان می‌رساند و دامنه کارهای آینده را ارائه می‌دهد.

۲- پیاده‌سازی روش پیشنهادی

در این پژوهش، مجموعه توییت‌های فارسی در مورد نرخ دلار در بازه زمانی سه ماهه تابستان سال ۱۳۹۷ شامل ماه اول ۱۹۲۸۵ توییت و ماه دوم ۱۱۶۵۰ توییت و ماه سوم ۷۶۷۰ توییت برای تحلیل احساسات درنظر گرفته شده است.

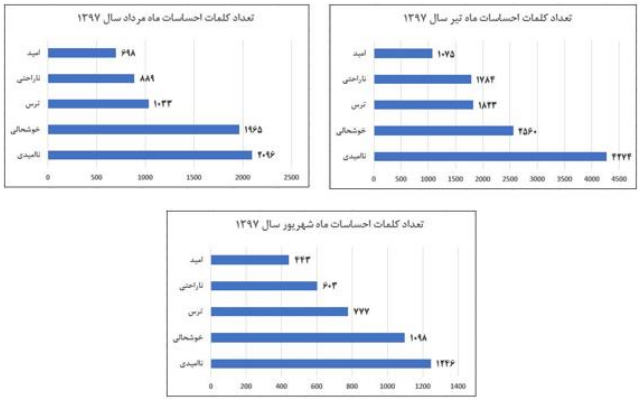
روش پیشنهادی برای بهبودی تحلیل احساسات یک فرهنگ‌نامه از کلمات احساسی (ناامیدی، خوشحالی، ترس، ناراحتی و امید) توییت‌های فارسی ایجاد کرده‌ایم. ابتدا توییت‌ها را از کلمات بی‌معنا پاک‌سازی کرده و براساس فرهنگ‌نامه احساسی موردنظر کلمات احساسی هر توییت را محاسبه می‌کنیم تا برچسب توییت‌ها را به درستی تشخیص دهیم و تحلیل احساسات در سطح جملات فارسی را بهبود می‌بخشیم. سپس برای متعادل‌سازی طبقه‌ها از روش نمونه‌گیری بیش از حد استفاده می‌کنیم و متون را به ماتریس عددی تبدیل کرده و با استفاده از الگوریتم‌های رگرسیون لجستیک (LR[15])، ماشین بردار پشتیبان(SVM[16])، درخت تصمیم (DT[17])، جنگل تصادفی (RF[18]) و روش جدید ترکیبی رگرسیون لجستیک و ماشین بردار پشتیبان (LR+SVM) و شبکه عصبی پیچشی (CNN[19]) و شبکه عصبی بازگشتی (LSTM[20]) روش پیشنهادی مورد ارزیابی قرار می‌گیرد. مراحل روش پیشنهادی را در شکل (۱) مشاهده می‌کنید.



شکل (۱): نمایی از روش پیشنهادی

۳- ارزیابی روش پیشنهادی

در این پژوهش ما یک فرهنگ‌نامه احساسی با عناوین ناامیدی، خوشحالی، ترس، ناراحتی و امید که هر کدام شامل کلمات مرتبط به آن‌ها است ایجاد کرده ایم و طبق این فرهنگ‌نامه با شمارش کلمات احساسی هر توییت برچسب احساسی آن را تعیین و مشخص کرده‌ایم. در نهایت تعداد هر کلمه احساسی برای هر سه ماه تابستان در شکل (۲) مشاهده می‌کنید. در نتیجه ناامیدی در ماه تیر با مقدار ۴۲۷۴ و در ماه مرداد با مقدار ۲۰۹۶ و در ماه شهریور با مقدار ۱۲۴۶ جهت نرخ دلار بیشترین تاثیر را روی افراد داشته است و امیدواری در ماه تیر با مقدار ۱۰۷۵ و در ماه مرداد با مقدار ۶۹۸ و در ماه شهریور با مقدار ۴۴۳ کمترین تاثیر را داشته است.



شکل (۲): تعداد کلمات احساسی در سه ماه تابستان سال ۱۳۹۷

همچنین در این پژوهش برای ارزیابی روش پیشنهادی از پنج الگوریتم یادگیری ماشین از جمله رگرسیون لجستیک (LR)، ماشین بردار پشتیبان(SVM)، درخت تصمیم (DT)، جنگل تصادفی (RF) و

روش جدید ترکیبی رگرسیون لجستیک و ماشین بردار پشتیبان (LR+SVM) و از دو الگوریتم یادگیری عمیق از جمله شبکه عصبی پیچشی (CNN) و شبکه عصبی بازگشتی (LSTM) دقت آن را در سه ماه تابستان در شکل (۳) بیان شده است.

همان‌طور که در شکل (۳) مشاهده می‌کنید در ماه تیر روش ترکیبی LR+SVM با مقدار ۹۹ درصد بیشترین دقت و LR و LSTM با مقدار ۹۶ درصد کمترین دقت و سایر الگوریتم‌ها از جمله SVM و DT و RF و CNN با مقدار ۹۸ درصد بهترین دقت را دارند. در ماه مرداد روش ترکیبی LR+SVM با مقدار ۹۷ درصد بیشترین دقت و LSTM با مقدار ۸۰ درصد کمترین دقت و سایر الگوریتم‌ها از جمله SVM و RF با مقدار ۹۶ درصد و DT با مقدار ۹۵ درصد و LR با مقدار ۹۲ درصد و CNN با مقدار ۹۰ درصد می‌باشند. در ماه شهریور روش ترکیبی LR+SVM با مقدار ۹۷ درصد بیشترین دقت و LSTM با مقدار ۷۱ درصد کمترین دقت و سایر الگوریتم‌ها از جمله DT و RF با مقدار ۹۶ درصد و SVM با مقدار ۹۴ درصد و LR با مقدار ۹۲ درصد و CNN با مقدار ۸۹ درصد می‌باشند.

شکل (۳): نتایج دقت الگوریتم‌های یادگیری ماشین سه ماه تابستان سال ۱۳۹۷

۴- نتیجه‌گیری

در این پژوهش سعی کردیم با روش ارائه شده و لغت‌نامه موردنظر، احساس هر یک از توییت‌های استخراجی سه ماه تابستان سال ۱۳۹۷ مربوط به نرخ دلار را پیدا کنیم. علاوه بر این یک روش ترکیبی از رگرسیون لجستیک و ماشین بردار پشتیبان ارائه دادیم. توییت‌ها ابتدا از کلمات اضافی پاک‌سازی می‌شوند. سپس تحت یک طبقه‌بندی احساسی شامل احساسات ناامیدی، خوشحالی، ترس، ناراحتی و امید لغت‌نامه‌ای ایجاد می‌کنیم و طبق لغت‌نامه احساس هر توییت را می‌یابیم. در چارچوب پیشنهادی، الگوریتم‌های رگرسیون لجستیک، ماشین بردار پشتیبان، درخت تصمیم و جنگل تصادفی و روش جدید ترکیبی رگرسیون لجیستیک و ماشین بردار پشتیبان و دو الگوریتم یادگیری عمیق از جمله الگوریتم‌های شبکه عصبی پیچشی و شبکه عصبی بازگشتی برای ارزیابی تحلیل احساسات مورد استفاده قرار می‌گیرد.

در نهایت به این نتیجه رسیدیم که احساسات مردم در مورد تغییر نرخ دلار در بازه زمانی سه ماهه تابستان سال ۱۳۹۷ بیشتر به سمت ناامیدی گرایش داشته است. همچنین روش ترکیبی ارائه شده برای ماه اول تابستان با مقدار ۹۹ درصد و دو ماه دیگر تابستان با مقدار ۹۷ درصد بیشترین دقت را نسبت به دیگر الگوریتم‌ها بهترین عملکرد را داشته است. همچنین الگوریتم‌های ماشین بردار پشتیبان(SVM) و درخت تصمیم (DT) و جنگل تصادفی (RF) و شبکه عصبی پیچشی (CNN) با مقدار ۹۸ درصد و رگرسیون لجستیک (LR) و شبکه عصبی بازگشتی (LSTM) با مقدار ۹۶ درصد بهترین دقت را به خود اختصاص داده‌اند. در زبان فارسی مشکلاتی چون عامیانه بودن کلمات و رعایت نکردن فاصله بین کلمات و رعایت نکردن قاعده ساختار جمله و استفاده از شکلک‌های تصویری می‌باشد که می‌توان این مشکلات را در کارهای آینده برای بهبودی تحلیل احساسات رفع کرد. همچنین می‌توان لغت-نامه احساسی را گسترش داد.

مراجع

[1] Yue, L., et al., *A survey of sentiment analysis in social media*. Knowledge and Information Systems, 2019.
[2] Feldman, R., *Techniques and applications for sentiment analysis*. Communications of the ACM, 2013.
[3] Zachman, John A., *A Framework for Information Systems Architecture*, IBM Systems Journal, Vol. 26, No. 3, 1987.
[4] Jain, A.P., Dandannavar, P., *Application of machine learning techniques to sentiment analysis*. in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). IEEE, 2016.
[5] Go, A., Bhayani, R., Huang, L, .*Twitter sentiment classification using distant supervision*. CS224N project report, Stanford, 2009. 1(12): p. 2009.
[6] Nakov, P., et al., SemEval-2016 task 4: *Sentiment analysis in Twitter*. arXiv preprint arXiv:1912.01973, 2019.

زیرنویس‌ها

[1] Sentiment Analysis

[2] Opinion Mining

[3] Data Mining
[4] Twitter
[5] Precision
[6] Recall
[7] Accuracy
[8] Naïve Bayes
[9] Multinomial Naïve Bayes
[10] Distant supervision
[11] Noisy labels
[12] Unigram
[13] Bigrams
[14] Part-Of-Speech
[15] Logistic Regression
[16] Support Vector Machine
[17] Diction Tree
[18] Random Forest
[19] Convolutional Neural Networks
[20] Long short-term memory