Tabriz Journal of Electrical Engineering (TJEE), vol. 54, no. 4, December 2024 Serial no. 110
DOI: 10.22034/tjee.2024.58494.4724

Using generative adversarial networks to increase the classification efficiency of imbalanced user reviews

Bahareh Javid, Hoda Mashayekhi*

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran[1, 2]
E-mails: bahareh.javid@shahroodut.ac.ir; hmashayekhi@shahroodut.ac.ir
* corresponding author

Short Abstract
Text generation methods use artificial intelligence to automatically generate natural language texts. One of the uses of text generation is in text classification. Many real-world problems are related to imbalanced textual data, which can reduce classification efficiency. One approach to solving the imbalanced data problem is the minority class oversampling. Due to the progress of generative adversarial networks (GAN) in data generation, these networks can be used to generate text samples in oversampling. Generating text using GANs is a complex problem due to the discrete nature of text. Despite their potential, the use of these networks in solving the problem of imbalanced textual data has rarely been investigated. This article examines the effect of using the SentiGAN network to solve the problem of imbalanced user reviews with the aim of improving the classification efficiency. To evaluate the proposed method, before and after oversampling with traditional, recent and SentiGAN methods, four classification algorithms were implemented on the data and evaluation criteria were calculated. It was observed that oversampling with the help of SentiGAN has increased the accuracy, precision, specificity and f_score of zero class compared to the situation where the data is imbalanced or even is oversampled by the other methods.

Keywords
Generative adversarial networks (GAN), imbalanced text classification, oversampling, imbalanced text, classification.

1- Short Introduction

In human activities, classification is one of the most used decision-making tasks. When dealing with real data, some conditions such as missing, imbalanced, and/or unlabeled data may reduce the accuracy and efficiency of the classifier. The imbalanced data problem in classification means that the number of instances in one (majority) class is much more than the number of instances in another (minority) class. Oversampling is a common data-driven preprocessing method for dealing with imbalanced data. Generating text by adversarial generative networks can use as an oversampling method to solve imbalanced text problem. Generating text by GANs is much more complex, because these models have not reached maturity in the scope of text generation.

2- Proposed Work and Methodology

In this article, it is proposed to use SentiGAN to solve the problem of imbalanced user reviews. Since GAN networks have the ability to generate data very similar to the original data, learn the internal representation of data, and learn disordered and complex distributions, they can be a good tool for oversampling. In order to evaluate the proposed method, first some classification algorithms were applied on imbalanced data and evaluation metrics were calculated. Then the data were balanced by traditional and recent oversampling methods (including SMOTE and ADASYN and a recent method). After that, the classification algorithms were applied to the data and the evaluation metrics were calculated. In the other part, the data were balanced with the help of SentiGAN network and classification algorithms were applied on them and evaluation metrics were calculated. At the end, the results of three cases were compared. In this we used Yelp database which includes customer reviews about restaurants. As a summary based on the observations of this research, it can be said that oversampling generally reduces the value of the recall compared to the case where the data are imbalanced. (In classification with random forest and naive Bayes algorithms, oversampling with SMOTE method increases the recall compared to the case where the data is imbalanced. Also, no change of recall can be seen in the classification with XGBoost.). On the other hand, oversampling with traditional and recent methods does not change the value of the precision (a slight decrease in the value of the precision in the logistic regression and XGBoost algorithm), but oversampling with the help of SentiGAN increases the value of the precision compared to that the data are imbalanced. Oversaming increases the Specificity compared to the case where the data is imbalanced, and this increase is greater when the data is balanced with the help of SentiGAN. Regarding the NPV, it was observed that oversampling with the help of SentiGAN increased this measure compared to imbalanced and balanced data using traditional and recent methods (of course, this was not observed in the XGBoost and Naive Bayes algorithms). In the following, it was observed that oversampling by traditional and recent methods does not change the accuracy (a very small decrease in the accuracy in the logistic regression algorithm and in the naïve bayes algorithm when balancing with the recent oversampling method also a very small increase in the accuracy in the random forest algorithm when Balancing with the SMOTE method) and oversampling with the help of SentiGAN increases the accuracy compared to the case where the data is imbalanced.

3- Conclusion

In this research, the effect of using GANs in increasing the efficiency of classification on imbalanced user reviews was investigated. In general, data balancing using SentiGAN is useful for increasing the classification efficiency of imbalanced textual data and has better overall performance compared to traditional oversampling methods. As future works, it is possible to investigate the effect of multi-class GAN networks in solving the problem of text imbalance and improving the efficiency of text classification. Investigating the efficiency of classification in the use of GAN networks in cases where the data is unbalanced and has multiple labels, as well as the use of GAN networks in solving the problem of missing data can be the subject of future research.

4- References

[1] K. Wang and X. Wan, “Automatic generation of sentimental texts via mixture adversarial networks,” Artif. Intell., vol. 275, pp. 540–558, 2019.
[2] I. J. Goodfellow et al., “Generative adversarial nets,” Adv. Neural Inf. Process. Syst., vol. 3, no. January, pp. 2672–2680, 2014.
[3] L. Yu, W. Zhang, J. Wang, and Y. Yu, “SeqGAN: Sequence generative adversarial nets with policy gradient,” 31st AAAI Conf. Artif. Intell. AAAI 2017, pp. 2852–2858, 2017.

استفاده از شبکه‌های مولد تخاصمی در افزایش کارایی دسته بندی نظرات نامتعادل کاربران

بهاره جاوید
دانشجوی دکتری، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، شاهرود، ایران

هدی مشایخی
دانشیار، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، شاهرود، ایران

چکیده
روش‌های تولید متن برای تولید خودکار متون زبان طبیعی از هوش مصنوعی استفاده می‌کنند. یکی از کاربردهای تولید متن در دسته‌بندی متن است. بسیاری از مسائل دنیای واقعی با داده‌های متنی نامتعادل در ارتباط هستند که می‌تواند کارایی دسته‌بندی را کاهش دهد. یک رویکرد حل مشکل داده‌های نامتعادل، بیش‌نمونه‌برداری از کلاس اقلیت است. با توجه به پیشرفت شبکه‌های مولد تخاصمی (GAN) در تولید داده، می‌توان از این شبکه‌ها برای تولید نمونه‌های متنی در بیش‌نمونه‌برداری استفاده کرد. تولید متن به کمک شبکه‌های مولد تخاصمی به دلیل ماهیت گسسته متن مسئله‌ای پیچیده است. علیرغم پتانسیل آن‌ها، استفاده این شبکه‌ها در حل مشکل داده‌های متنی نامتعادل به ندرت مورد بررسی قرار گرفته است. این مقاله به بررسی تاثیر استفاده از شبکه‌ی SentiGAN بر حل مشکل عدم تعادل نظرات کاربران با هدف بهبود کارایی دسته‌بندی می‌پردازد. بعد از ارائه روش پیشنهادی و چارچوب ارزیابی، چهار الگوریتم دسته‌بندی بر روی داده‌ها اجرا شده و معیارهای ارزیابی مختلف پیش و پس از بیش‌نمونه‌برداری محاسبه و تحلیل شده‌اند. هم‌چنین نتایج با روش‌های بیش‌نمونه‌برداری سنتی و اخیر مقایسه شده است. بیش‌نمونه‌برداری با روش پیشنهادی باعث افزایش معیارهای صحت، دقت و تشخیص‌پذیری، و امتیاز اف دسته‌بندی داده‌های اقلیت نسبت به داده‌های نامتعادل و همچنین در مقایسه با روش‌های دیگر بیش‌نمونه‌برداری می‌شود.

کلمات کلیدی
شبکه‌های مولد تخاصمی (GAN)، دسته‌بندی متون نامتعادل، بیش‌نمونه‌برداری، متن نامتعادل، دسته بندی.

نام نویسنده مسئول: هدی مشایخی
ایمیل نویسنده مسئول: hmashayekhi@shahroodut.ac.ir

تاریخ ارسال مقاله: ۱۴۰۲/۰۶/۲۷
تاریخ(های) اصلاح مقاله: ۱۴۰۲/۱۰/۰۲
تاریخ پذیرش مقاله: ۱۴۰۲/۱۱/۱۵

۱- مقدمه

دسته‌بندی یکی از پراستفاده‌ترین وظایف تصمیم‌سازی است. یک مسئله دسته‌بندی زمانی اتفاق می‌افتد که نیاز باشد یک شیء براساس تعدادی از ویژگی‌های مشاهده شده از آن شیء به یک گروه از پیش‌تعریف‌شده یا کلاس نسبت داده شود. بسیاری از مسائل در علم، صنعت، پزشکی و تجارت می‌توانند به عنوان یک مسئله دسته‌بندی در نظر گرفته شوند [۱]. در دسته‌بندی داده-های واقعی شرایطی پیش می‌آید که این شرایط باعث کاهش دقت و کارایی دسته‌بند می‌شوند. از این شرایط می‌توان به وجود داده‌های مفقود (Missing)، عدم تعادل کلاس داده‌ها و وجود داده‌های بدون برچسب اشاره کرد [۲-۴]. مسأله عدم تعادل کلاس زمانی وجود دارد که نمونه‌های یکی از کلاس‌ها (کلاس اکثریت) بسیار بیشتر از کلاس دیگر (کلاس اقلیت) است. این مسئله فقط در داده‌ها با دو کلاس اتفاق نمی‌افتد بلکه ممکن است در داده‌های چند کلاسه نیز اتفاق بیفتد [۵]. برخی از کاربردها که از مجموعه داده‌های نامتعادل استفاده می‌کنند عبارت هستند از تشخیص‌های پزشکی [۶]، شناسایی مشتری‌های غیر قابل اعتماد در ارتباط از راه دور، تشخیص نشت نفت در تصاویر راداری ماهواره-ای، یادگیری تلفظ کلمات، تشخیص تماس‌های تلفنی جعلی و بازیابی اطلاعات [۵]. همچنین در برخی مسائل با داده‌های متنی می‌توان به تشخیص هرزنامه [۷]، تجزیه و تحلیل احساسات جملات [۷]، تشخیص کارت‌های اعتباری تقلبی از معتبر [۸]، تشخیص و مدیریت ترافیک جاده‌ای براساس داده‌های متنی شبکه اجتماعی [۹]، مدیریت منابع انسانی [۳]، انجمن‌های بحث و پاسخ [۱۰] و غیره اشاره کرد. هنگامی که داده‌ها نامتعادل هستند، هزینه گم شدن داده‌ها در کلاس اقلیت از هزینه گم شدن داده‌ها در کلاس اکثریت بالاتر است، اکثر سیستم‌های یادگیری برای مقابله با اختلاف زیاد بین تعداد نمونه‌های متعلق به هر کلاس آمادگی ندارند و الگوریتم‌های دسته‌بندی عملکرد مناسبی ندارند [۱۱]. در هنگام آموزش یک روش دسته‌بندی استاندارد، کلاس اقلیت کمتر در به حداقل رساندن تابع هدف کمک می‌کند. از طرفی تمایز ایجاد کردن بین نمونه‌های کلاس اقلیت و نمونه‌های نویز اغلب کار سختی است. نکته مهم این است که در بسیاری از کاربردها هزینه اشتباه دسته‌بندی کردن داده‌های کلاس اقلیت بسیار بالاتر از هزینه اشتباه دسته‌بندی کردن داده‌های کلاس اکثریت است [۱۲]. روش‌های متعددی برای برخورد با داده‌های نامتعادل به منظور افزایش کارایی و دقت دسته‌بند پیشنهاد شده است. یکی از این روش‌ها بیش‌نمونه‌برداری (Oversampling) است. ایده اصلی بیش‌نمونه‌برداری افزایش اندازه کلاس اقلیت است تا کلاس‌های متعادل به دست بیاید. بیش‌نمونه‌برداری یک رویکرد رایج بر پایه داده در برخورد با داده‌های نامتعادل است که جزء روش‌های پیش‌پردازشی محسوب می‌شود. از مزایای این روش می‌توان به از دست ندادن اطلاعات برخلاف روش کم نمونه برداری (Undersampling) اشاره کرد [۱۳].

مدل‌سازی مولد یک روش یادگیری بدون نظارت در یادگیری ماشین است که مدل برای یادگیری الگوهای موجود در داده آموزش می‌بینند. بعد از آموزش

و با استفاده از دانش به دست آمده مدل قادر خواهد بود داده‌های جدیدی تولید کند که با داده‌های آموزشی مرتبط هستند [۱۴]. تولید داده توسط شبکه‌های عصبی عمیق و شبکه‌های مولد عمیق که توزیع حقیقی داده‌ها را یاد می‌گیرند، در سال‌های اخیر پیشرفت چشمگیری داشته است [۱۵]. امروزه با حجم زیادی از داده‌های متنی روبرو هستیم. ایمیل‌ها، متن‌های خبری، گزارش‌های علمی و تکنیکی، کتاب‌ها، گزارشات قضایی، نظامی و پزشکی نمونه‌هایی از این داده‌ها هستند [۱۶]. تولید متن زیربخشی از پردازش زبان طبیعی است که از هوش مصنوعی برای تولید خودکار متون استفاده می‌کند و در زمینه‌هایی مانند ترجمه ماشینی [۱۷]، داستان‌گویی خودکار [۱۸]، خلاصه‌سازی متن [۱۹]، تولید کلمات کلیدی [۲۰]، سیستم‌های پرسش و پاسخ [۲۱]، تولید اخبار مصنوعی [۲۲]، تولید شعر [۲۳] و غیره کاربرد دارد. تولید متن توسط شبکه‌های مولد عمیق در زمینه متن بسیار پیچیده‌تر از تولید تصویر است، چون شبکه‌های مولد عمیق در تولید متن به بلوغ نرسیده و تولید متن در آن‌ها با مشکلاتی روبرو است. ساختار گسسته متن انتشار گرادیان از تفکیک‌کننده به مولد را در آموزش استاندارد شبکه‌های مولد تخاصمی[1] (GAN) غیر ممکن کرده است. دو مشکل دیگر در شبکه‌های GAN که در برخورد با داده‌های متنی تشدید می‌شوند عبارت هستند از بی‌ثباتی آموزش[2] و افت حالت[3]. در افت حالت، یک حالت خاص در مجموعه آموزشی به ندرت توسط مولد ایجاد می‌شود. در تولید متن این یک مشکل قابل توجه به دلیل پیچیدگی متن است. بی‌ثباتی آموزش نیز یک مشکل است زیرا متن به صورت خودکار تولید می‌شود و خطای تفکیک‌کننده پس از تولید جملات کامل مشاهده می‌شود. این مشکل در تولید جملات طولانی‌تر پیچیده‌تر می‌شود [۲۴].

در این پژوهش به بررسی تاثیر تولید متن توسط شبکه‌های GAN بر افزایش کارایی دسته‌بندی نظرات نامتعادل کاربران پرداخته می‌شود. دسته‌بندی نظرات در حوزه عقیده‌کاوی یا تحلیل احساسات قرار می‌گیرد. در این حوزه سعی بر تشخیص و بیان احساسات، نظرات، رفتار و تحلیل افراد مختلف در مورد یک موجودیت و ویژگی های آن است [۲۵]. این پژوهش از شبکه GAN برای بیش‌نمونه‌برداری استفاده می‌کند تا نظرات کاربران متعادل شود و کارایی دسته‌بندی افزایش یابد. به این منظور شبکه GAN متنی با نام SentiGAN با استفاده از داده‌های کلاس اقلیت آموزش داده می‌شود. با اضافه کردن نمونه‌های تولید شده توسط مدل آموزش‌دیده به مجموعه داده نامتعادل، تعداد نمونه‌های کلاس اقلیت و اکثریت برابر می‌شود و در واقع بیش‌نمونه‌برداری انجام خواهد شد. به منظور ارزیابی روش پیشنهادی، داده‌ها با روش‌های سنتی و اخیر بیش-نمونه‌برداری هم متعادل شدند و با استفاده از چهار الگوریتم دسته‌بندی متفاوت داده‌ها در کلیه حالات (از جمله داده‌های نامتعادل و داده‌های متعادل شده با روش‌های مختلف) دسته‌بندی شدند و معیارهار ارزیابی محاسبه و نتایج مقایسه شدند. استفاده از روش‌های مولد برای حل مشکل عدم تعادل متن در مقالات بسیار کم دیده شده است و پژوهش‌های کمی در این زمینه انجام شده‌اند. از دلایل این موضوع می‌توان به پیچیدگی تولید متن با روش‌های مولد عمیق اشاره کرد. در ادامه مقاله در بخش دوم ادبیات و پیشینه پژوهش بیان می‌شود. در بخش سوم روش پیشنهادی تشریح شده و در بخش چهارم آزمایش‌ها و نتایج ذکر می‌شوند. در انتها نتیجه‌گیری بیان خواهد شد.

۲- ادبیات و پیشینه پژوهش

در ادامه پیشینه‌ای راجع به تولید متن بیان می‌شود. سپس با توجه به اینکه تمرکز این مقاله بر شبکه‌های GAN می‌باشد، به معرفی این شبکه‌ها پرداخته می‌شود. پس از آن پیشینه ای راجع به کاربرد شبکه‌های GAN در رفع مشکل عدم تعادل ذکر خواهد شد.

۲-۱- تولید متن

تولید زبان طبیعی (NLG) به عنوان یک رویکرد سیستماتیک برای تولید متن قابل فهم توسط انسان براساس داده‌های غیرمتنی یا بازنمایی‌های معنادار تعریف می‌شود [۲۶]. انواع مختلفی از سیستم‌های تولید زبان طبیعی وجود دارد. ساده‌ترین آن‌ها متن آماده (Canned text) و پس از آن سیستم‌های پر کردن قالب (Template filling) هستند و در انتها سیستم‌های پیشرفته ای هستند که سازگار با تغییرات واقع‌بینانه در اطلاعات یک دامنه می‌باشند. سیستم‌های پیشرفته باید در انتخاب محتوا، گزینش واژگان، تجمیع ساختار جمله و ساختار گفتمان تصمیم‌گیری داشته باشند [۲۷].

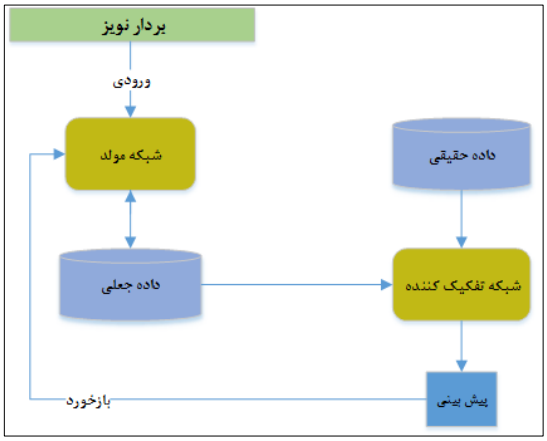
برخی از موفقیت‌های اولیه در زمینه تولید زبان، ساخت سیستم‌هایی مانند Eliza [۲۸] و PARRY [۲۹] است. این سیستم‌ها زبان را از طریق مجموعه‌ای از قوانین تولید می‌کنند. با این حال چنین سیستم‌هایی مبتنی بر قوانین بیش از حد محدود و شکننده هستند و نمی‌توان آن‌ها را به راحتی برای تولید مجموعه‌ای از پاسخ‌ها تعمیم داد. سایر تکنیک‌های سنتی تولید زبان طبیعی متن را از داده‌های ساخت‌یافته یا از پایگاه‌های دانش تولید می‌کنند. برخی از نمونه‌ها سیستم‌های مبتنی بر دامنه هستند که گزارش‌های هواشناسی [۳۰] یا ورزشی [۳۱] را تولید می‌کنند [۳۲]. زمینه سیستم‌های تولید متن از رویکرد-های قدیمی به رویکردهای آماری تغییر یافته است که در آن تمرکز بر استفاده از الگوهای موجود در داده متنی و ساخت مدل‌ها به منظور پیش‌بینی مناسب براساس متن دیده شده می‌باشد. میکولوف و همکاران [۳۳] اظهار داشتند که پیشرفت چشم‌گیری در استفاده از رویکردهای آماری برای مدل‌سازی زبان وجود نداشته است. این مشاهدات منجر به آزمایش وی روی استفاده از شبکه-های عصبی بازگشتی و دستیابی به نتایج پیشرفته‌ای شد که باعث تبدیل شبکه-های عصبی به الگوی انتخابی برای مدل‌سازی داده‌های متوالی مانند متن شوند. شبکه‌های عصبی به طبقه‌ای از مدل‌های یادگیری ماشینی تعلق دارند که قادر به شناسایی الگوهای موجود در متن هستند و ویژگی‌هایی را شناسایی می‌کنند که به حل مشکلات مختلف مربوط به بینایی کامپیوتر (Computer vision)، تشخیص اشیاء (Object recognition)، زیرنویس کردن تصویر (Image captioning) و تشخیص گفتار (Speech recognition) کمک می‌کند [۳۴]. از سوی دیگر، در دسترس بودن تعداد زیادی از منابع محاسباتی اتفاق خوبی برای ظهور شبکه‌های عصبی بود. از جمله الگوریتم‌های عمیقی که در تولید متن به کار می‌روند، می‌توان به شبکه‌های عصبی بازگشتی (RNN)، شبکه‌های LSTM، شبکه‌های عصبی بازگشتی دوطرفه (BRNN) اشاره کرد. در سال‌های اخیر از شبکه‌های VAE و شبکه‌های GAN نیز در تولید متن استفاده شده است [۱۵]. از جمله مزایای شبکه‌های GAN می‌توان به تولید داده‌هایی که بسیار شبیه داده‌های اصلی هستند، یادگیری بازنمایی داخلی داده‌ها و توان یادگیری توزیع-های نامرتب و پیچیده و قابلیت کار با داده‌های بدون برچسب اشاره کرد. در این پژوهش سعی داریم با استفاده از شبکه GAN تولید متن انجام دهیم و با استفاده از این روش کارایی دسته‌بندی متن را در حالتی که دادها نامتعادل هستند، افزایش دهیم.

۲-۲- شبکه‌های مولد تخاصمی

این شبکه‌ها برای بار اول توسط گودفیلو و همکاران [۳۵] معرفی شدند. یک شبکه GAN از دو شبکه عصبی مصنوعی تشکیل شده که با که با یکدیگر به رقابت می پردازند: مولد (Generator) و تفکیک‌کننده (Discriminator). بخش

[1] Generative adversarial networks
[2] Training instability
[3] Mode dropping

اول یک نمونه جدید داده تولید می‌کند و بخش دوم اعتبار آن را ارزیابی می‌کند [۳۶]. تفکیک‌کننده مسئول ارزیابی کیفیت داده‌های ایجاد شده توسط مولد است و به عنوان ورودی، نمونه‌های داده را از مجموعه داده اصلی و یا داده‌های تولید شده توسط مولد می‌گیرد و سعی می‌کند منبع نمونه را پیش‌بینی کند. مولد یاد می‌گیرد یک فضای نهان[۴] را به توزیع داده‌ها نگاشت کند و قصد دارد که بازتولید انجام دهد. بنابراین زمانی که با یک بردار نویز از یک فضای نهان تغذیه می‌شود، نمونه‌ای از توزیع تخمین زده شده را پیش‌بینی می‌کند. مولد به وسیله تفکیک کننده ارزیابی می‌شود، به این معنی که هدف آن ایجاد نمونه داده‌هایی است که مشابه نمونه‌های موجود در مجموعه داده اصلی هستند. با آموزش همزمان هر دو شبکه، آن‌ها در رقابت با یکدیگر بهتر می‌شوند. از این رو نام شبکه‌های مولد تخاصمی برای آن‌ها انتخاب شده است. تفکیک کننده سعی می‌کند در تشخیص داده‌های جعلی و واقعی بهتر شود و مولد به دنبال تولید داده‌هایی است که به تدریج به داده‌های واقعی نزدیک‌تر شوند [۳۶]. شکل ۱ این عملکرد را نشان می‌دهد.



شکل ۱ - چگونگی آموزش همزمان مولد و تفکیک کننده در GAN

این مسئله می‌تواند به‌صورت زیر فرموله شود:

فرض کنید مولد G و تفکیک‌کننده را با D نشان می‌دهیم. توزیع مولد روی داده x یا P_G به صورت یک تابع مشتق‌پذیر $G(z;\theta_g)$ مدل شده است که می‌تواند به وسیله یک شبکه عصبی با پارامترهای θ_g و نویز ورودی z پیاده‌سازی شود. تفکیک‌کننده $D(x;\theta_d)$ نیز توسط یک شبکه عصبی با پارامترهای θ_d پیاده‌سازی می‌شود، نمونه x را به عنوان ورودی می‌گیرد و یک عدد اسکالر به عنوان خروجی می‌دهد که نشان‌دهنده‌ی احتمال این‌که x از داده‌ها به جای P_G می‌آید، می‌باشد. هدف آموزش GAN یادگیری توزیع مولد $P_G(x)$ است که با توزیع داده حقیقی $P_{data}(x)$ مطابقت داشته باشد [۳۵]. برای تحقق این هدف، یک تابع بازی minimax به صورت رابطه ۱ پیشنهاد می‌گردد:

$$\min_G \max_D V(D.G) = E_{x\sim P_{data}}[log\, D(x)] + E_{z\sim noise}[log\,(1-D(G(Z)))] \quad (۱)$$

در فرایند آموزش، تفکیک‌کننده قصد دارد $D(G(z))$ را به عدد صفر برساند و $D(x)$ را به یک برساند. در حالی که مولد سعی می کند $D(G(z))$ را به عدد یک نزدیک کند تا احتمال اشتباه افتادن را به حداکثر برساند. این ساختار شبکه با یک بازی minimax دو بازیکنه مطابقت دارد [۳۷]. به‌طور مشخص اگر اندازه دسته متغیرهای نویز و نمونه‌های آموزش m باشد، نیاز داریم تفکیک‌کننده را با بالابردن گرادیان تصادفی آن بروزرسانی نماییم:

$$\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}[\log D(x^{(i)}) + \log(1-D(G(z^{(i)})))] \quad (۲)$$

به طور عکس مولد با کاهش گرادیان تصادفی آپدیت می‌شود:

$$\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m} log(1-D(G(z^{(i)}))) \quad (۳)$$

پس از آموزش مناسب، راه‌حل ایده‌آل برای این بازی minimax تعادل نش[۵] است. الگوریتم کلی GAN در شکل ۲ آمده است:

الگوریتم ۱: آموزش یک مدل GAN

برای تعداد تکرارهای آموزش انجام بده:

برای k گام انجام بده:

دسته کوچکی شامل m نمونه نویز $z^{(1)}, \dots, z^{(m)}$ از نویز پیشین $p_g(z)$ نمونه برداری کن.

دسته کوچکی شامل m نمونه $x^{(1)}, \dots, x^{(m)}$ از احتمال پیشین $p_{data}(z)$ نمونه برداری کن.

تفکیک کننده را با افزایش گرادیان تصادفی آن بروز کن.

$$\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}[\log D(x^{(i)}) + \log(1-D(G(z^{(i)})))]$$

پایان حلقه.

دسته کوچکی شامل m نمونه نویز $z^{(1)}, \dots, z^{(m)}$ از نویز پیشین $p_g(z)$ نمونه برداری کن.

مولد را با کاهش گرادیان تصادفی آن بروز کن .

$$\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}\log(1-D(G(z^{(i)})))$$

پایان حلقه.

به روز رسانی های مبتنی بر گرادیان می توانند از هر قانون استاندارد یادگیری مبتنی بر گرادیان استفاده کنند.

شکل ۲- الگوریتم کلی GAN [۳۵]

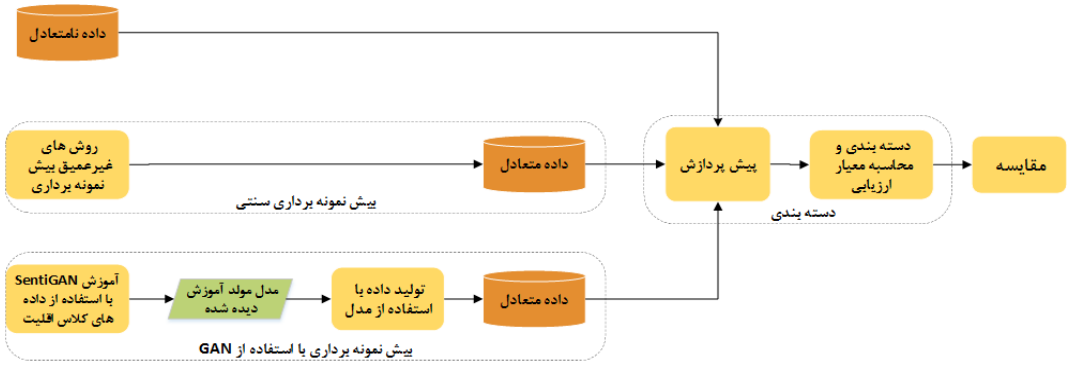
در این پژوهش ما قصد داریم تاثیر استفاده از شبکه‌های GAN در افزایش کارایی دسته‌بندی متن را بررسی نماییم، بنابراین از شبکه SentiGAN [۳۸] استفاده کردیم که یکی از شبکه‌هایی است که برای تولید متن طراحی شده است. همان‌طور که در مقدمه ذکر شد، تولید متن با شبکه‌های GAN به دلیل ماهیت گسسته متن با محدودیت‌هایی مواجه است. برای حل این مشکلات و محدودیت‌ها، محققان مدل‌های GAN متنی را معرفی کردند. وانگ و وان در سال ۲۰۱۹ شبکه SentiGAN [۳۸] را معرفی کردند که هدف آن تولید متون از کلاس‌های مجزا به صورت بدون نظارت است. SentiGAN دارای چندین مولد هرکدام برای یک کلاس جداگانه و یک تفکیک کننده چند کلاسه است. هر مولد مجبور می‌شود نمونه‌های متنوعی از یک کلاس خاص را با استفاده از یک هدف مبتنی بر جریمه تولید کند. در ادامه پیشینه‌ای از استفاده شبکه‌های GAN در بهبود دسته‌بندی ذکر شده است.

۳-۲- استفاده از شبکه‌های مولد تخاصمی در بهبود دسته‌بندی

با توجه به تحقیقات زیادی که بر روی شبکه‌های GAN انجام شده است، در سال‌های اخیر برخی از این شبکه‌ها برای حل مشکل عدم تعادل استفاده شده‌اند. دوزاس و همکاران [۳۹] از شبکه‌های GAN شرطی برای تولید نمونه هایی از کلاس اقلیت استفاده کردند که می‌تواند توزیع مجموعه داده (تصویری) را بهبود دهد. مائو و همکاران [۴۰] از شبکه‌های GAN برای تولید نمونه‌هایی از کلاس اقلیت در هنگام انجام آزمایشات تشخیص خطا استفاده کردند. سپس این نمونه‌ها را به مجموعه آموزش اضافه کردند. فیور و همکاران [۴۱] با مجموعه

[4] Latent space

[5] Nash equilibrium



شکل ۳ - مراحل انجام تحقیق

داده‌های نامتعادل در زمان تشخیص خطای کارت‌های اعتباری روبرو شدند، آن‌ها از GAN به عنوان یک روش بیش نمونه‌برداری برای تولید نمونه‌هایی از کلاس اقلیت استفاده کردند و باعث افزایش کارایی دسته‌بند شد. به منظور رفع مشکل دسته‌بندی در حالت عدم تعادل در متن لئو و همکاران [۴۲] در مقاله ای از معماری تولید متن SeqGAN [۴۳] استفاده کردند. در این مقاله ابتدا دسته‌بندی انجام شد. سپس معیارهای F1-score و G-mean برای کلاس اقلیت محاسبه شدند. سپس به وسیله SeqGAN متن‌هایی از کلاس اقلیت ساخته شد و به مجموعه داده‌های کلاس اقلیت اضافه شد. نتایج آزمایشات نشان داد که بیش نمونه‌برداری با استفاده از SeqGAN بر بقیه روش‌های نمونه‌برداری غلبه می‌کند و نشان می‌دهد این روش علاوه بر افزایش کارایی کلی دسته‌بند، به استخراج ویژگی‌های بیشتر از کلاس اقلیت نیز کمک می‌کند.

در پژوهش‌های انجام شده موارد کمی وجود دارند که بر روی بهبود کارایی رده‌بندی متن نامتعادل با استفاده از شبکه‌های مولد کار کرده باشند.

۳- روش پیشنهادی

همان‌طور که ذکر شد در دسته‌بندی داده‌های واقعی شرایطی مانند وجود داده‌های مفقود، عدم تعادل کلاس داده‌ها و وجود داده‌های بدون برچسب باعث کاهش دقت و کارایی دسته‌بند می‌شوند. روش‌های زیادی برای افزایش دقت و کارایی رده‌بندی در هر کدام از این شرایط وجود دارد. در این پژوهش تمرکز بر افزایش کارایی دسته‌بندی متون (نظرات کاربران) نامتعادل با استفاده از شبکه‌های GAN می‌باشد. بسیاری از داده‌های دنیای واقعی نامتعادل هستند. به‌علت عدم تعادل دسته‌بند نمی‌تواند با دقت خوبی دسته نمونه‌ها را پیش‌بینی نماید و این امر باعث بروز خطا می‌شود. در برخورد با داده‌های نامتعادل یک رویکرد بیش‌نمونه‌برداری است و به معنای افزایش داده‌های کلاس اقلیت به منظور تعادل داده‌های کلاس‌ها است. یک روش برای بیش‌نمونه‌برداری استفاده از شبکه‌های GAN برای تولید داده از کلاس اقلیت است. از آنجایی که شبکه‌های GAN قابلیت تولید داده‌های بسیار شبیه به داده اصلی، یادگیری بازنمایی داخلی داده‌ها و یادگیری توزیع‌های نامرتب و پیچیده را دارند می‌توانند ابزار خوبی برای بیش‌نمونه‌برداری باشند. از طرفی در پژوهش‌های کمی کاربرد شبکه‌های GAN در بهبود کارایی دسته‌بندی متن نامتعادل دیده شده است. رویکرد پیشنهادی برای استفاده از شبکه‌های GAN در حل مشکل عدم تعادل و افزایش کارایی دسته‌بندی در شکل ۳ نمایش داده می‌شود. به‌منظور ارزیابی روش پیشنهادی ابتدا برخی از الگوریتم‌های دسته‌بندی بر روی داده‌های نامتعادل اعمال شدند و معیارهای ارزیابی محاسبه شدند. سپس داده‌ها به وسیله روش‌های سنتی بیش-نمونه‌برداری (شامل SMOTE و ADASYN) و یک روش اخیر بیش‌نمونه‌برداری ارائه شده در [۴۴] متعادل شدند. پس از آن الگوریتم‌های دسته بندی بر روی داده‌ها اعمال شده و معیارهای ارزیابی محاسبه شدند. در بخش دیگر داده‌ها با کمک شبکه SentiGAN متعادل شدند و الگوریتم‌های دسته‌بندی روی آن‌ها اعمال گردید و معیارهای ارزیابی محاسبه شدند. در انتها نتایج سه حالت با یکدیگر مقایسه شدند. در ادامه هر یک از این مراحل شرح داده خواهد شد.

۳-۱- بیش‌نمونه‌برداری به کمک روش‌های سنتی

به‌منظور متعادل سازی داده‌ها به وسیله روش‌های سنتی دو روش SMOTE [۴۵] و ADASYN [۴۶] استفاده شدند. SMOTE یک روش موفق و پر ارجاع-ترین روش بیش‌نمونه‌برداری برای حل مشکل عدم تعادل کلاس‌هاست [۷]. ایده اصلی SMOTE، بیش‌نمونه‌برداری از کلاس اقلیت با تولید تصادفی اشیاء مصنوعی از این کلاس است [۴۷]. برای تولید یک شی مصنوعی، تفاوت ویژگی بین یک شی از کلاس اقلیت و یکی از k نزدیک‌ترین همسایه‌های انتخابی تصادفی آن محاسبه می‌شود. سپس برای هر ویژگی، یک انحراف (Offset) با ضرب این تفاوت در یک عدد تصادفی در بازه واحد ایجاد می‌شود. در نهایت، یک شی مصنوعی با افزودن افستها به شی اصلی انتخاب شده از کلاس اقلیت تولید می شود [۴۷]. روش ADASYN دو هدف دارد که عبارت هستند از: اول، به‌منظور کاهش عدم تعادل بین کلاس اقلیت و کلاس اکثریت، نمونه‌های مصنوعی را از طریق درونیابی خطی بین نمونه‌های اقلیت تولید می‌کند. هدف دوم که ADASYN را با توجه به SMOTE متفاوت می‌کند اینکه داده های تولید شده به طور تطبیقی مرز تصمیم را تغییر می دهد. این کار را با افزودن داده‌ها در ناحیه کلاس اقلیت که یادگیری آن دشوار است در مقایسه با داده‌های کلاس اقلیت که یادگیری آن آسان است، از طریق توزیع چگالی انجام می‌دهد. ADASYN قصد دارد به داده های کلاس اقلیت که یادگیری آن‌ها دشوار است، وزن بیشتری بدهد [۴۸].

در این بخش پس از متعادل کردن داده‌ها با استفاده از روش‌های سنتی بیش‌نمونه‌برداری، الگوریتم‌های طبقه‌بندی اعمال شده و معیارهای ارزیابی محاسبه می‌شوند.

۳-۲- بیش‌نمونه‌برداری با استفاده از شبکه مولد تخاصمی

برای استفاده از GAN متنی در حل مشکل عدم تعادل کلاس‌ها، ابتدا GAN با نمونه های کلاس اقلیت آموزش داده می‌شود. شبکه GAN توزیع داده‌های کلاس اقلیت را یاد می‌گیرد و مدل آموزش دیده می‌تواند نمونه‌هایی تولید کند که به کلاس اقلیت شباهت دارند. سپس نمونه‌های تولید شده به کمک مدل به مجموعه داده نامتعادل اضافه می‌شود تا تعداد نمونه‌های کلاس اقلیت و اکثریت یکسان شود. پس از متعادل کردن داده‌ها با روش های ذکر شده، مجدداً الگوریتم های طبقه بندی اعمال شده و معیارهای ارزیابی محاسبه می‌شوند. در ادامه ساختار شبکه GAN استفاده شده با نام SentiGAN توضیح داده شده است.

وانگ و وان [۳۸] مدلی را برای تولید متن با برچسب های مختلف پیشنهاد کردند. مدل آن‌ها شامل چندین مولد و یک تفکیک کننده چند کلاسه است. در

مدل آن‌ها چندین مولد به طور هم‌زمان با هدف تولید متن از کلاس‌های مختلف آموزش داده می‌شود. یک هدف مبتنی بر جریمه (penalty-based) برای مولدها پیشنهاد شده است، که هر یک آن‌ها را مجبور می‌کند نمونه‌های مختلفی از یک برچسب خاص را تولید کنند. استفاده از مولدهای متعدد و تفکیک کننده چندکلاسه نیز می‌تواند هر مولد را مجبور کند تا بر تولید دقیق متن خود با یک کلاس خاص تمرکز کند.

همانطور که ذکر شد برای تولید متن از k کلاس، مدل SentiGAN دارای k مولد $\{G_i(x_t|x_{0:t-1};\theta_g^i)\}_{i=1}^{i=k}$ و یک تفکیک‌کننده $D(x;\theta_d)$ که θ_g^i و θ_d به ترتیب پارامترهای i امین مولد و تفکیک‌کننده هستند. به منظور آموزش مولد، مسئله تولید متن به صورت یک فرآیند تصمیم‌سازی ترتیبی فرموله می‌شود. این روش استفاده می‌شود تا مسئله ناتوانی گرادیان برای انتشار به عقب به سمت مدل مولد در مواجهه با داده‌های گسسته را حل کند. در هر نقطه زمانی t، یک مولد G_i برای تولید یک رشته $x_{0:t}=\{x_0.x_1.\dots.x_t\}$ آموزش می‌بیند که x_t یک کلمه از دیکشنری C را نشان می‌دهد. تابع زیان (Loss) مبتنی بر جریمه (Penalty) مولد به صورت زیر تعریف می‌شود:

$$L(x)=\sum_{t=1}^{|x|-1} G_i\left(x_t|x_{0:t-1};\theta_g^i\right)\cdot V_{D_i}^{G_i}(x_{0:t-1}\oplus x_t) \qquad (۴)$$

که $G_i\left(x_t|x_{0:t-1};\theta_g^i\right)$ احتمال انتخاب کلمه t ام با توجه به کلمات $x_{0:t-1}$ که قبلا تولید شده‌اند را نشان می‌دهد. $\oplus$ عملگر الحاق است و $V_{D_i}^{G_i}(x_{0:t-1}\oplus x_t)$ جریمه برای رشته $x_{0:t-1}\oplus x_t$ محاسبه شده توسط تفکیک کننده است. از آنجایی که تفکیک کننده تنها قادر به داوری بر اساس رشته کامل است، برای نمونه برداری $|x|-t$ توکن ناشناخته آخر، جستجوی مونت کارلو با سیاست roll-out (G_i) اعمال می‌شود. در ادامه طریقه محاسبه تابع جریمه برای مولد i ام مشاهده می‌شود:

$$V_{D_i}^{G_i}(x_{1:t-1}\oplus x_t)=\begin{cases}\frac{1}{N}\sum_{n=1}^{N}\left(1-D_i\left(x_{0:t}\oplus x_{t+1:|x|}^n;\theta_d\right)\right) & t<|x|\\ 1-D_i(x_{0:t};\theta_d) & t=|x|\end{cases} \qquad (۵)$$

احتمال جمله بدست آمده بوسیله تفکیک کننده با $D_i\left(x_{0:t}\oplus x_{t+1:|x|}^n;\theta_d\right)$ نشان داده می‌شود که $x_{0:t}\oplus x_{t+1:|x|}^n$ متن واقعی از کلاس i ام است. هدف i امین مولد $G_i\left(x_t|x_{t-1};\theta_g^i\right)$ کمینه کردن مقدار کلی تعریف شده بر اساس پنالتی است:

$$J_{G_i}(\theta_g^i)=\mathbb{E}_{x\sim P_{g_i}}[L(X)]=\mathbb{E}_{x\sim P_{g_i}}[\sum_{t=1}^{|x|-1} G_i\left(x_t|x_{0:t-1};\theta_g^i\right)\cdot V_{D_i}^{G_i}(x_{0:t-1}\oplus x_t)] \qquad (۶)$$

که $x_t\in C$ است. مولد در این مدل یک LSTM تک لایه است که به‌عنوان خروجی کلمه t ام را براساس توزیع تولید می‌کند:

$$p_t=softmax\left(LSTM_{\theta_g}(h_{t-1}.x_{t-1})\right).\sum_{t\in C}p_t(x_t)=1 \qquad (۷)$$

که θ_g پارامترهای مولد $LSTM_{\theta_g}$ و h_t حالت نهان[6] در زمان t می‌باشد. در ادامه نحوه آموزش تفکیک‌کننده در مدل SentiGAN بیان می‌شود. با در نظر گرفتن مجموعه ای از k مولد، تفکیک‌کننده یک توزیع احتمال Softmax بر روی $k+1$ کلاس ایجاد می کند. احتمال تعلق به متن واقعی از کلاس i ام با D_i (امتیاز اندیس i ام) و احتمال تولید نمونه به صورت مصنوعی با D_{k+1} (امتیاز اندیس $k+1$ ام) نمایش داده می‌شود. تابع هدف تفکیک‌کننده که باید کمینه شود به‌صورت زیر است:

$$J_D(\theta_d)=-\mathbb{E}_{x\sim P_g}\log D_{k+1}(x;\theta_d)-\sum_{i=1}^{k}\mathbb{E}_{x\sim P_{r_i}}\log D_i(x;\theta_d) \qquad (۸)$$

که P_g متن تولید شده توسط همه مولدهاست، P_{r_i} متن واقعی با برچسب i است و $D_i(x;\theta_d)$ امتیاز اندیس i ام $D(x;\theta_d)$ می‌باشد. تفکیک‌کننده لایه‌ای از CNN است که چندین فیلتر دارد.

۳-۳- دسته‌بندی

روش‌های دسته‌بندی بر روی داده نامتعادل و متعادل‌شده به منظور مقایسه کارایی اعمال شدند. در این مطالعه از چهار دسته‌بند جنگل تصادفی (Random Forest) [۴۹]، رگرسیون لاجستیک (Logistic Regression) [۵۰]، بیز ساده (Naïve Bayes) [۵۱] و XGBoost [۵۲] برای دسته‌بندی متون استفاده شده است. رگرسیون لاجستیک یک دسته‌بند دودویی است که از تابع سیگموئید برای دسته‌بندی استفاده می‌کند. یک بردار ویژگی را می‌گیرد و به عنوان خروجی، یک احتمال طبقه‌بندی را برمی‌گرداند. جنگل تصادفی یک دسته‌بند جمعی[7] است که از چندین درخت تصمیم‌گیری تشکیل شده است و نتیجه دسته‌بندی نهایی آن بر اساس رأی اکثریت در میان دسته‌بندهای فردی است. دسته‌بند بیز ساده یک الگوریتم دسته بندی آماری رایج است که برای دسته‌بندی داده‌ها از قضیه بیز و فرض استقلال شرطی استفاده می‌کند. XGBoost یک الگوریتم یادگیری ماشین جمعی بر پایه درخت تصمیم است که از چارچوب gradient boosting استفاده می‌کند. دلیل انتخاب این الگوریتم‌های دسته‌بندی استفاده از آنها در تحقیقات مشابه مانند مقالات [۵۳] و [۵۴] و [۴۲] می‌باشد.

۴- آزمایشات

در این بخش آزمایشاتی که برای ارزیابی روش پیشنهادی انجام شدند ذکر می‌شوند. در بخش ۴-۱ مجموعه داده و در بخش ۴-۲ تنظیمات استفاده شده در آزمایشات بیان می‌شود. معیارهای ارزیابی در بخش ۴-۳ معرفی می‌شوند و نتایج آزمایشات در بخش ۴-۴ ذکر خواهند شد.

۴-۱- مجموعه داده

در این پژوهش از مجموعه داده Yelp[8] استفاده شد. مجموعه داده Yelp شامل نظرات مشتریان درباره رستوران‌ها است و در آن امتیازات اصلی مشتری به یک برچسب مثبت یا منفی باینری تبدیل می‌شود. در مجموعه داده آموزشی از مجموع ۲۴۰۰۰۰ هزار نمونه موجود، ۲۱۶۰۰۰ نمونه از کلاس مثبت یا یک و ۲۴۰۰۰ نمونه از کلاس منفی یا صفر در نظر گرفته شدند. مجموعه داده آزمایشی شامل ۱۶۰۰۰۰ نمونه می‌باشد که ۸۰۰۰۰ از کلاس منفی یا صفر و ۸۰۰۰۰ از کلاس مثبت یا یک است. اطلاعات مجموعه داده استفاده شده به صورت خلاصه در جدول ۱ نمایش داده شده است.

جدول ۱- اطلاعات مجموعه داده استفاده شده

داده آزمایشی		داده آموزشی	
۱۶۰۰۰۰		۲۴۰۰۰۰	
کلاس منفی (صفر)	کلاس مثبت (یک)	کلاس منفی (صفر)	کلاس مثبت (یک)
۸۰۰۰۰	۸۰۰۰۰	۲۴۰۰۰	۲۱۶۰۰۰

[6] Hidden state
[7] Ensemble classifier
[8] "Yelp Dataset." https://www.yelp.com/dataset

۴-۲- تنظیمات آزمایش‌ها

کلیه آزمایشات روی سیستمی با سیستم عامل Windows 10 64-bits، پردازشگر Intel(R) Xeon(R) CPU E5-2695 v3 @ 2.30GHz 2.30GH (2 processors)، حافظه رم ۲۵۶ گیگابایتی و پردازنده گرافیکی NVIDIA Quadro M4000 با حافظه رم ۸ گیگابایتی اجرا شده است. پیاده‌سازی SentiGAN از طریق کد قابل دسترس مخزن Github [9] انجام شده است. این پیاده سازی نیاز به داده برای پیش‌آموزش دارد، بنابراین نیمی از داده‌های آموزشی (داده‌های کلاس اقلیت) به عنوان داده پیش‌آموزش در نظر گرفته شدند. مدل با ۱۰۰۰ تکرار (epoch) آموزش داده شده است.

۴-۳- معیارهای ارزیابی

در این پژوهش برای بررسی کارایی دسته‌بندی قبل و بعد از بیش‌نمونه‌برداری از معیارهای زیر استفاده شده است:

- **صحت**: معیار صحت (Accuracy) نشان می‌دهد چند نمونه داده به درستی پیش‌بینی شده‌اند و از تقسیم تعداد نمونه‌های مثبت درست پیش‌بینی شده یا مثبت صحیح (True Positive یا TP) و تعداد نمونه‌های منفی درست پیش‌بینی شده یا منفی صحیح (True Negative یا TN) بر مجموع TP و TN و تعداد نمونه‌های منفی نادرست پیش بینی شده یا مثبت کاذب (False Positive یا FP) و تعداد نمونه‌های مثبت نادرست پیش‌بینی شده یا منفی کاذب (False Negative یا FN) به‌دست می‌آید:

$$\text{صحت} = \frac{\text{تعداد پیش‌بینی های درست}}{\text{تعداد کل پیش‌بینی ها}} \quad (۹)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **یادآوری**: معیار یادآوری (Recall) نسبت نمونه‌های مثبت درست پیش‌بینی شده در بین تمام نمونه‌های مثبت در مجموعه آزمایشی را محاسبه می‌کند. برای محاسبه آن، TP بر مجموع TP و FN تقسیم می‌شود:

$$Recall = \frac{TP}{TP + FN} \quad (۱۰)$$

- **دقت:** دقت (Precision) تعداد نمونه‌های واقعی مثبت در میان تمام نمونه‌های پیش‌بینی شده مثبت در مجموعه آزمایشی را نشان می‌دهد. برای محاسبه آن TP بر مجموع TP و FP تقسیم می‌شود:

$$Precision = \frac{TP}{TP + FP} \quad (۱۱)$$

- **تشخیص‌پذیری**: معیار تشخیص‌پذیری (Specificity) محاسبه می‌کند که چند نمونه منفی به درستی در بین تمام نمونه‌های منفی در مجموعه آزمایشی پیش‌بینی شده است. برای محاسبه آن TN بر مجموع TN و FP تقسیم می‌شود:

$$Specificity = \frac{TN}{TN + FP} \quad (۱۲)$$

- **امتیاز اف**: میانگین هارمونیک (Harmonic) معیارهای یادآوری و دقت امتیاز اف (F-measure) نامیده می‌شود و به صورت زیر محاسبه می‌گردد:

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (۱۳)$$

- NPV (negative predictive value): NPV در واقع دقت کلاس صفر یا منفی است و عبارت است از تعداد نمونه‌های واقعی منفی در بین تمام نمونه‌های پیش‌بینی شده منفی در مجموعه آزمایشی.

$$NPV = \frac{TN}{TN + FN} \quad (۱۴)$$

- امتیاز اف کلاس صفر: در واقع میانگین هارمونیک معیارهای تشخیص‌پذیری و NPV است.

$$F\_score\ of\ zero\ class = 2 * \frac{NPV * \text{Specificity}}{NPV + \text{Specificity}} \quad (۱۵)$$

۴-۴- نتایج آزمایشات

در این بخش نتایج آزمایشات ذکر می‌شود. ابتدا معیارهای ارزیابی کارایی دسته‌بندی گزارش خواهد شد و سپس تحلیل نتایج بیان می‌شود.

کارایی دسته‌بندی

معیارهای کارایی دسته‌بندی پس از اعمال الگوریتم دسته‌بندی بر روی هر یک از داده‌های نامتعادل، متعادل شده با SMOTE، ADASYN و متعادل شده با روش اخیر بیش‌نمونه‌برداری ارائه شده در [۴۴] همچنین متعادل شده با SentiGAN محاسبه شده است. جدول ۲ مقادیر معیارهای ارزیابی برای سه الگوریتم دسته‌بندی اجرا شده بر روی پایگاه داده ذکر شده را نمایش می‌دهد. برای چهار الگوریتم رگرسیون لاجستیک، جنگل تصادفی و بیز ساده و XGBoost مقادیر معیارهای صحت، تشخیص‌پذیری، یادآوری، دقت، NPV، امتیاز اف و امتیاز اف کلاس صفر برای داده‌ها در هر یک از حالات ذکر شده محاسبه شده است. بیشترین مقدار در هر حالت با خط در زیر عدد مشخص شده است.

تحلیل نتایج

در ادامه تغییرات معیارهای کارایی دسته‌بندی قبل و بعد از متعادل کردن داده‌ها مورد بررسی قرار می‌گیرد. در این بخش معیارهای ارزیابی برای دسته‌بند رگرسیون لاجستیک بررسی می‌شوند. به منظور بررسی بهتر معیارها، ماتریس در هم ریختگی (Confusion) مربوط به دسته‌بند رگرسیون لاجستیک در هر یک از حالات نامتعادل و متعادل شده در جدول ۲ نمایش داده شده است.

جدول ۲- جدول در هم ریختگی مربوط به دسته‌بند رگرسیون لاجستیک در هر یک از حالات متعادل و نامتعادل

	TN	FP	FN	TP	SUM
داده نامتعادل	۱۶۱۶۰	۶۳۸۴۰	۱۵۵۹۸	۶۴۴۰۲	۱۶۰۰۰۰
داده متعادل‌شده با روش SMOTE	۲۵۶۶۴	۵۴۳۳۶	۲۶۷۰۴	۵۳۲۹۶	۱۶۰۰۰۰
داده متعادل‌شده با روش ADASYN	۲۶۰۳۷	۵۳۹۶۳	۲۷۲۰۳	۵۲۷۹۷	۱۶۰۰۰۰
داده متعادل شده با کمک مقاله [۴۴]	۲۷۶۶۵	۵۲۳۳۵	۲۸۱۸۰	۵۱۸۲۰	۱۶۰۰۰۰
داده متعادل شده با کمک SentiGAN	۵۷۷۶۳	۲۲۲۳۷	۵۰۵۳۷	۲۹۴۶۳	۱۶۰۰۰۰

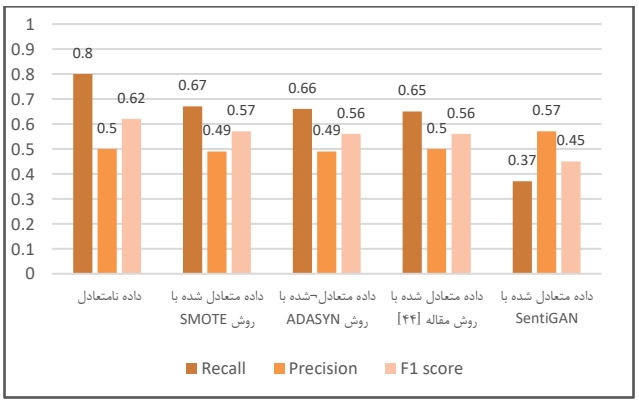
در ابتدا به بررسی معیارهایی می‌پردازیم که شرایط کلاس اکثریت یعنی کلاس با برچسب یک و احساسات مثبت را بهتر نمایش می‌دهند. شکل ۴ تغییرات معیارهای یادآوری، دقت و امتیاز اف را برای دسته‌بند رگرسیون لاجستیک قبل و بعد از متعادل کردن داده‌ها نمایش می‌دهد.

[9] https://github.com/Nrgeup/SentiGAN

مشاهده می‌شود متعادل کردن داده‌ها باعث کاهش مقدار معیار یادآوری نسبت به حالتی که داده‌ها نامتعادل هستند، می‌شود. این کاهش مقدار معیار یادآوری در حالتی که داده‌ها با استفاده از SentiGAN متعادل شده‌اند، بیشتر است. معیار یادآوری نسبت نمونه‌های TP را در همه نمونه‌های مثبت محاسبه می‌کند. طبق جدول ۳ مشاهده می‌شود بیش‌نمونه‌برداری داده‌ها باعث کاهش کاهش TP دسته‌بند نسبت به حالتی که داده‌ها نامتعادل هستند، می‌شود. همچنین متعادل کردن باعث افزایش FN می‌شود. این دو باعث می‌شوند که یادآوری کاهش یابد. این نتیجه قابل انتظار است زیرا نمونه‌های تولید شده از کلاس اقلیت دسته‌بند را از بیش‌برازش (Overfitting) روی کلاس اکثریت منع می‌کنند. در واقع کاهش TP به خاطر این است که قبل از متعادل کردن، تعداد نمونه‌های کلاس مثبت نسبت به تعداد نمونه‌های کلاس منفی بیشتر هستند و مدل می‌تواند خیلی خوب آن‌ها را شناسایی کند.

از طرفی در شکل ۴ مشاهده می‌شود بیش‌نمونه‌برداری با کمک SentiGAN باعث افزایش مقدار معیار دقت نسبت به حالتی که داده‌ها نامتعادل هستند، می‌شود. با توجه به جدول ۴ آموزش دسته‌بند بر روی داده‌های متعادل شده باعث کاهش TP و FP نسبت به آموزش بر روی داده‌های نامتعادل می‌شود. کاهش TP همانطور که ذکر شد مورد انتظار است چون نمونه‌های تولید شده از کلاس اقلیت، دسته‌بند را از بیش‌برازش بر روی داده‌های کلاس اکثریت منع می‌کنند و این موضوع باعث افزایش خطای دسته‌بند در تشخیص کلاس اکثریت می‌شود. کاهش FP همچنین می‌تواند با کاهش بیش‌برازش بر روی داده‌های کلاس اکثریت توضیح داده شود، که باعث می‌شود دسته‌بند دو کلاس را بهتر تشخیص داده و نمونه‌های منفی (کلاس اقلیت) را به کلاس صحیح آن‌ها تخصیص دهد.

معیار امتیاز اف میانگین هارمونیک مقادیر معیارهای یادآوری و دقت است و با توجه به مقادیر این دو معیار در هر یک از حالات نامتعادل و متعادل شده مشاهده می‌شود متعادل کردن داده‌ها باعث کاهش معیار امتیاز اف نسبت به حالتی که داده‌ها نامتعادل هستند می‌شود. این کاهش در حالتی که داده‌ها با استفاده از SentiGAN متعادل شده‌اند، بیشتر دیده می‌شود. دلیل این امر کاهش بیشتر معیار یادآوری در متعادل کردن با استفاده از SentiGAN می‌باشد.



شکل ۴ -تغییرات معیارهای یادآوری، دقت و امتیاز اف برای دسته‌بند رگرسیون لاجستیک قبل و بعد از متعادل کردن داده‌ها

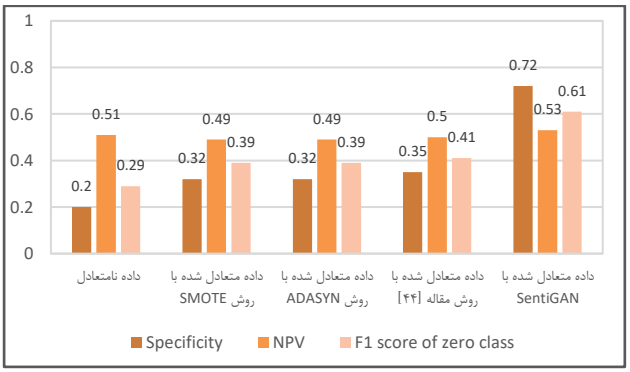
با توجه به اینکه کلاس اقلیت، کلاس صفر (کلاس احساسات منفی) در نظر گرفته شده است، بعضی از معیارهای ارزیابی که نشان‌دهنده بهتر شرایط کلاس اقلیت و تاثیر بیش‌نمونه‌برداری از آن هستند، نیز مورد بررسی قرار می‌گیرند. این معیارها عبارت هستند از تشخیص‌پذیری یا Recall کلاس صفر، NPV یا دقت کلاس صفر و امتیاز اف کلاس صفر. شکل ۵ تغییرات این معیارها را برای دسته‌بند رگرسیون لاجستیک قبل و بعد از متعادل کردن داده‌ها نمایش می‌دهد.

جدول ۳ - مقادیر معیارهای ارزیابی برای چهار الگوریتم دسته‌بندی اجرا شده

مجموعه داده YELP		داده نامتعادل	داده متعادل شده با روش SMOTE [۴۵]	داده متعادل‌شده با روش ADASYN [۴۶]	داده متعادل شده با روش مقاله [۴۴]	داده متعادل شده با SentiGAN
رگرسیون لاجستیک	صحت	۰.۵۰	۰.۴۹	۰.۴۹	۰.۵۰	۰.۵۴
	تشخیص‌پذیری	۰.۲۰	۰.۳۲	۰.۳۲	۰.۳۵	۰.۷۲
	یادآوری	۰.۸۰	۰.۶۷	۰.۶۶	۰.۶۵	۰.۳۷
	NPV	۰.۵۱	۰.۴۹	۰.۴۹	۰.۵۰	۰.۵۳
	دقت	۰.۵۰	۰.۴۹	۰.۴۹	۰.۵۰	۰.۵۷
	امتیاز اف کلاس صفر	۰.۲۹	۰.۳۹	۰.۳۹	۰.۴۱	۰.۶۱
	امتیاز اف	۰.۶۲	۰.۵۷	۰.۵۶	۰.۵۶	۰.۴۵
جنگل تصادفی	صحت	۰.۵۱	۰.۵۲	۰.۵۱	۰.۵۱	۰.۵۵
	تشخیص‌پذیری	۰.۳۴	۰.۳۴	۰.۳۷	۰.۳۳	۰.۵۵
	یادآوری	۰.۶۸	۰.۷۰	۰.۶۵	۰.۶۹	۰.۵۵
	NPV	۰.۵۲	۰.۵۳	۰.۵۱	۰.۵۲	۰.۵۵
	دقت	۰.۵۱	۰.۵۱	۰.۵۱	۰.۵۱	۰.۵۵
	امتیاز اف کلاس صفر	۰.۴۱	۰.۴۱	۰.۴۳	۰.۴۰	۰.۵۵
	امتیاز اف	۰.۵۸	۰.۵۹	۰.۵۷	۰.۵۹	۰.۵۵
بیز ساده	صحت	۰.۵۲	۰.۵۲	۰.۵۲	۰.۵۱	۰.۵۵
	تشخیص‌پذیری	۰.۲۱	۰.۱۹	۰.۲۲	۰.۴۶	۰.۶۹
	یادآوری	۰.۸۳	۰.۸۵	۰.۸۳	۰.۵۷	۰.۴۲
	NPV	۰.۵۵	۰.۵۵	۰.۵۶	۰.۵۱	۰.۵۴
	دقت	۰.۵۱	۰.۵۱	۰.۵۱	۰.۵۱	۰.۵۷
	امتیاز اف کلاس صفر	۰.۳۰	۰.۲۸	۰.۳۱	۰.۴۸	۰.۶۱
	امتیاز اف	۰.۶۳	۰.۶۴	۰.۶۳	۰.۵۴	۰.۴۸
XGBoost	صحت	۰.۵۱	۰.۴۹	۰.۴۹	۰.۵۱	۰.۵۲
	تشخیص‌پذیری	۰.۰۷	۰.۷۰	۰.۷۲	۰.۰۷	۰.۹۴
	یادآوری	۰.۹۵	۰.۲۸	۰.۲۵	۰.۹۵	۰.۱۰
	NPV	۰.۵۷	۰.۴۹	۰.۴۹	۰.۵۵	۰.۵۱
	دقت	۰.۵۰	۰.۴۸	۰.۴۸	۰.۵۰	۰.۶۵
	امتیاز اف کلاس صفر	۰.۱۳	۰.۵۸	۰.۵۸	۰.۱۲	۰.۶۶
	امتیاز اف	۰.۶۶	۰.۳۵	۰.۳۳	۰.۶۶	۰.۱۸

همان‌طور که مشاهده می‌شود بیش‌نمونه‌برداری باعث افزایش معیار تشخیص-پذیری نسبت به حالتی که داده‌ها نامتعادل هستند، می‌شود و این افزایش هنگامی که داده‌ها با کمک SentiGAN متعادل می‌شوند، بیشتر است. با توجه به جدول ۳ دلیل افزایش معیار تشخیص‌پذیری بعد از متعادل کردن داده‌ها، افزایش TN و کاهش FP است. در واقع بیش‌نمونه‌برداری دسته‌بند را قادر به تشخیص بهتر داده‌های کلاس اقلیت کرده است. با توجه به اینکه کلاس اقلیت در این پژوهش کلاس منفی بوده است، معیار تشخیص‌پذیری اطلاعات بهتری راجع به تاثیر متعادل کردن داده‌ها در اختیار قرار می‌دهد. از طرفی در مورد معیار NPV، بیش‌نمونه‌برداری به کمک روش‌های سنتی و اخیر باعث کاهش یا

عدم تغییر این معیار نسبت به حالتی که داده‌ها نامتعادل هستند، شده است اما هنگامی که داده‌ها با استفاده از SentiGAN متعادل می‌شوند، این معیار نسبت به حالتی که داده‌های نامتعادل هستند یا با روش‌های سنتی و اخیر متعادل شده‌اند، افزایش می‌یابد. با توجه به جدول ۳ دلیل این امر، افزایش TN نسبت به حالتی که داده‌ها نامتعادل‌هستند یا با روش‌های سنتی و اخیر متعادل شده‌اند، است. همانطور که قبلا ذکر شد بیش‌نمونه‌برداری باعث شده دسته‌بند نمونه‌های کلاس اقلیت را بهتر تشخیص دهد. در اینجا مشاهده می‌شود FN نیز نسبت به حالتی که داده‌ها نامتعادل هستند، افزایش یافته است. دلیل این امر این است که بعد از متعادل کردن داده‌ها مدل نمونه‌های کلاس منفی یا صفر را نیز به اندازه نمونه‌های کلاس مثبت مشاهده می‌کند و تعداد بیشتری از نمونه‌های کلاس مثبت را اشتباه به کلاس منفی نسبت می‌دهد.



شکل ۵ - تغییرات معیارهای تشخیص‌پذیری، NPV و امتیاز اف کلاس صفر برای دسته‌بند رگرسیون لاجستیک قبل و بعد از متعادل کردن داده‌ها

در انتها با توجه به جدول ۲ مشاهده می‌شود بیش‌نمونه‌برداری با کمک SentiGAN باعث افزایش معیار صحت نسبت به حالتی که داده‌ها نامتعادل هستند، می‌شود. با توجه به جدول ۳ و موارد مطرح شده در بالا دلیل افزایش مقدار معیار صحت کاهش TP، افزایش TN، کاهش FP و افزایش FN است که این امر در اثر زیاد شدن تعداد داده‌های کلاس اقلیت (منفی) و جلوگیری از بیش برازش مدل بر روی کلاس اکثریت است.

به عنوان یک جمع‌بندی براساس مشاهدات این پژوهش می‌توان بیان کرد به صورت کلی بیش‌نمونه‌برداری باعث کاهش مقدار معیار یادآوری نسبت به حالتی که داده‌ها نامتعادل هستند، می‌شود (در دسته‌بندی با الگوریتم‌های جنگل تصادفی و بیز ساده بیش‌نمونه‌برداری با روش SMOTE باعث افزایش معیار یادآوری نسبت به حالتی که داده‌ها نامتعادل هستند، می‌شود. همچنین در دسته‌بندی با XGBoost عدم تغییر یادآوری دیده می‌شود). از طرفی بیش‌نمونه-برداری با روش‌های سنتی و اخیر باعث تغییر نکردن مقدار معیار دقت (کاهش جزئی مقدار معیار دقت در الگوریتم رگرسیون لاجستیک و XGBoost) می‌شود اما بیش‌نمونه‌برداری با کمک SentiGAN باعث افزایش مقدار معیار دقت نسبت به حالتی که داده‌ها نامتعادل هستند و بقیه حالات، می‌شود. در ادامه مشاهده شد بیش‌نمونه‌برداری باعث افزایش معیار تشخیص‌پذیری نسبت به حالتی که داده‌ها نامتعادل هستند، می‌شود و این افزایش هنگامی که داده‌ها با کمک SentiGAN متعادل می‌شوند بیشتر است. در مورد معیار NPV مشاهده شد بیش‌نمونه برداری به کمک SentiGAN باعث افزایش این معیار نسبت به حالات داده نامتعادل و متعادل شده به کمک روش‌های سنتی و اخیر گردید (البته این مورد در الگوریتم XGBoost و بیز ساده مشاهده نشد). در مورد معیار صحت، بیش‌نمونه‌برداری به وسیله روش‌های سنتی باعث تغییر نکردن معیار صحت (کاهش خیلی کم معیار صحت در الگوریتم رگرسیون لاجستیک و افزایش خیلی کم معیار صحت در الگوریتم جنگل تصادفی هنگام متعادل کردن با روش SMOTE و کاهش خیلی کم صحت در الگوریتم بیز ساده هنگام متعادل کردن با روش اخیر بیش‌نمونه‌برداری) و بیش‌نمونه‌برداری با کمک SentiGAN باعث افزایش معیار صحت نسبت به حالتی که داده‌ها نامتعادل هستند، می‌شود.

۵- نتیجه‌گیری

در این پژوهش به بررسی تاثیر استفاده از شبکه‌های GAN در افزایش کارایی دسته‌بندی نظرات نامتعادل کاربران پرداخته شد. به این منظور مدل تولید متن SentiGAN انتخاب شد و چارچوبی برای حل مشکل عدم تعادل متن ارائه شد. به منظور ارزیابی روش پیشنهادی، بعد از متعادل کردن داده‌ها با هر کدام از روش‌های سنتی و اخیر بیش‌نمونه‌برداری و SentiGAN، چهار الگوریتم دسته‌بندی متفاوت بر روی داده‌ها اعمال شدند و معیارهای ارزیابی دسته‌بندی محاسبه شدند. به طور کلی، متعادل کردن داده ها با استفاده از شبکه SentiGAN برای افزایش کارایی دسته‌بندی داده‌های متنی نامتعادل مفید است و عملکرد کلی بهتری در مقایسه با روش های بیش‌نمونه‌برداری سنتی دارد. به عنوان کارهای آینده، می‌توان به بررسی تاثیر شبکه‌های GAN چندکلاسه در حل مشکل عدم تعادل متن و بهبود کارایی دسته‌بندی متن پرداخت. بررسی کارایی دسته‌بندی در استفاده از شبکه‌های GAN در مواردی که داده‌ها نامتعادل و دارای چند برچسب هستند و همچنین استفاده از شبکه‌های GAN در حل مشکل وجود داده‌های مفقود می‌توانند موضوع تحقیقات آینده باشند.

سپاس‌گزاری

نویسندگان از حمایت مالی وزارت علوم، تحقیقات و فناوری از این پژوهش در قالب کد اعتباری ۱۶-۰۰-۰۲-۰۰۰۴۵۵ قدردانی می‌نمایند.

مراجع

[1] G. P. Zhang, “Neural networks for classification: a survey”, IEEE Trans. Syst. Man, Cybern. Part C Appl. Rev., vol. 30, no. 4, pp. 451–462, 2000.

[2] T. R. Baitharu and S. K. Pani, “Effect of Missing Values on Data Classification Corresponding Author : Tapas Ranjan Baitharu,” *journals.co.za*, vol. 4, no. 2, pp. 311–316, 2013.

[3] C. Padurariu and M. E. Breaban, “Dealing with data imbalance in text classification”, in Procedia Computer Science, vol. 159, pp. 736–745, 2019.

[4] I. Glaser, S. Sadegharmaki, B. Komboz, and F. Matthes, “Data scarcity: Methods to improve the quality of text classification”, In ICPRAM, pp. 556-564. 2021.

[5] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Handling imbalanced datasets : A review”, GESTS international transactions on computer science and engineering , vol. 30, no. 1, pp. 25–36, 2006.

[6] Z. Xu, D. Shen, T. Nie, and Y. Kou, “A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data”, Journal of Biomedical Informatics, vol. 107, p. 103465, Jul. 2020.

[7] J. Tian, S. Chen, X. Zhang, and Z. Feng, “A graph-based measurement for text imbalance classification”, European Conference on Artificial Intelligence, pp. 2188–2195, 2020.

[8] H. He and E. A. Garcia, “Learning from imbalanced data”, IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, 2009.

[9] T. Raksachat and R. Chawuthai, “Improving a text classifier using text augmentation: road traffic content from Twitter”, In 2023 20th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) (pp. 1-4), 2023.

[10] A. Amin et al., “Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study”, IEEE Access, vol. 4, pp. 7940–7957, 2016.

[11] A. Sonak and R. A. Patankar, “A Survey on Methods to Handle Imbalance Dataset”, Int. J. Comput. Sci. Mob. Comput., vol. 4, no. 11, pp. 338–343, 2015.

[12] G. Douzas and F. Bacao, “Effective data generation for imbalanced learning using conditional generative adversarial networks”, Expert Systems with applications, vol. 91, no. January 2018, pp. 464–471, 2018.

[13] H. Kaur, H. S. Pannu, and A. K. Malhi, “A systematic review on imbalanced data challenges in machine learning: Applications and solutions”, ACM Computing Surveys (CSUR), vol. 52, no. 4, 2019.

[14] Available online at: https://www.section.io/engineering-education/beginners-intro-to-generative-modeling/#discriminative-and-generative-modeling.

[15] T. Iqbal and S. Qureshi, “The survey: Text generation models in deep learning”, Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 6, 2022.

[16] مرضیه رحیمی، عرفان جلیلی جلال، حسین رحیمی، « تولید کلمات کلیدی متون فارسی با استفاده از یادگیری انتقالی»، مجله مهندسی برق دانشگاه تبریز، جلد ۵۲، شماره ۲، صفحات ۱۲۳-۱۱۵، ۱۴۰۱.

[17] I. Rivera-Trigueros, “Machine translation systems and quality assessment: a systematic review”, Language Resources and Evaluation, vol. 56, no. 2, pp. 593–619, 2022.

[18] Y. Mori, H. Yamane, Y. Mukuta, and T. Harada, “Computational Storytelling and Emotions: A Survey”, arXiv (Cornell University), May 2022.

[19] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Automatic text summarization: A comprehensive survey”, Expert Systems with Applications, vol. 165. Pergamon, p. 113679, 2021.

[20] M. Scholz, C. Brenner, and O. Hinz, “AKEGIS: automatic keyword generation for sponsored search advertising in online retailing”, Decision Support Systems, vol. 119, pp. 96–106, 2019.

[21] B. Ojokoh and E. Adebisi, “A review of question answering systems”, Journal of Web Engineering, vol. 17, no. 8. pp. 717–758, 2019.

[22] K. Shu, Y. Li, K. Ding, and H. Liu, “Fact-Enhanced Synthetic News Generation”, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 15, no. 15, pp. 13825–13833, 2021.

[23] S. Talafha and B. Rekabdar, “Arabic Poem Generation Incorporating Deep Learning and Phonetic CNNsubword Embedding Models”, International Journal of Robotic Computing, pp. 64–91, 2019.

[24] W. Fedus, I. Goodfellow, and A. M. Dai, “MaskGaN: Better text generation via filling in the”, 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., 2018.

[25] سعید دهقانی اشکذری، ولی درهمی، علی محمد زارع بیدکی، محمداحسان بصیری، « عقیده‌کاوی در زبان فارسی مبتنی بر یادگیری انتقالی»، مجله مهندسی برق دانشگاه تبریز، جلد ۵۰، شماره ۳، صفحات ۱۲۲۴-۱۲۱۵، ۱۳۹۹.

[26] M. Wielgosz et al., “Evaluation and implementation of n-gram-based algorithm for fast text comparison”, Computing and Informatics, vol. 36, no. 4, pp. 887–907, 2017.

[27] J. G. Saliby, “Survey on Natural Language Generation”, International Journal of Trend in Scientific Research and Development, vol. Volume-3, no. Issue-3, pp. 618–622, 2019.

[28] J. Weizenbaum, “ELIZA-A computer program for the study of natural language communication between man and machine”, Communications of the ACM, vol. 9, no. 1, pp. 36–45, 1966.

[29] K. M. Colby, “Artificial paranoia: A computer simulation of paranoid processes”, Behavior Therapy, vol. 7, no. 1, p. 146, Jan. 1976.

[30] G. Angeli, P. Liang, and D. Klein, “A simple domain-independent probabilistic approach to generation”, in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 502-512. 2010.

[31] R. Barzilay and L. Lee, “Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization”, arXiv (Cornell University), pp. 113–120, 2004.

[32] S. Santhanam and S. Shaikh, “A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems - Past, Present and Future Directions.,” arXiv (Cornell University), 2019

[33] T. Mikolov, M. Karafiát, L. Burget, C. Jan, and S. Khudanpur, “Recurrent neural network based language model”, Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, vol. 2, pp. 1045–1048,2010.

[34] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks”, Advances in Neural Information Processing Systems, vol. 4, no. January, pp. 3104–3112, 2014.

[35] I. Goodfellow et al., “Generative Adversarial Nets”, Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680, 2014.

[36] F. H. K. dos S. Tanaka and C. Aranha, “Data Augmentation Using GANs”, *arXiv [cs.LG]*, 2019.

[37] Y. Zhang, “Deep Generative Model for Multi-Class Imbalanced Learning”, 2018..

[38] K. Wang and X. Wan, “Automatic generation of sentimental texts via mixture adversarial networks”, Artificial Intelligence, vol. 275, pp. 540–558, 2019.

[39] G. Douzas and F. Bacao, “Effective data generation for imbalanced learning using conditional generative adversarial networks”, Expert Systems with applications, vol. 91, pp. 464–471, 2018.

[40] W. Mao, Y. Liu, L. Ding, and Y. Li, “Imbalanced fault diagnosis of rolling bearing based on generative adversarial network: A comparative study”, IEEE Access, vol. 7, pp. 9515–9530, 2019.

[41] U. Fiore, A. De Santis, F. Perla, P. Zanetti, and F. Palmieri, “Using generative adversarial networks for improving classification effectiveness in credit card fraud detection”, *Information Sciences*, vol. 479, pp. 448–455, 2019.

[42] Y. Luo, H. Feng, X. Weng, K. Huang, and H. Zheng, “A novel oversampling method based on SeqGAN for imbalanced text classification”, 2019 IEEE International Conference on Big Data (Big Data), pp. 2891–2894, 2019.

[43] L. Yu, W. Zhang, J. Wang, and Y. Yu, “SeqGAN: Sequence generative adversarial nets with policy gradient”, 31st AAAI Conf. Artif. Intell. AAAI 2017, pp. 2852–2858, 2017.

[44] S. Bej, N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer, “LoRAS: an oversampling approach for imbalanced datasets”, Machine Learning, vol. 110, no. 2, pp. 279–301, 2021.

[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique”, Journal of artificial intelligence research , vol. 16, pp. 321–357, 2002.

[46] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”, in 2008 IEEE international joint conference on neural networks, pp. 1322–1328.

[47] F. Rodríguez-Torres, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, “An Oversampling Method for Class Imbalance Problems on Large Datasets”, Applied Sciences, vol. 12, no. 7, 2022.

[48] M. Torres-Vásquez, J. Hernández-Torruco, B. Hernández-Ocaña, and O. Chávez-Bosquez, “Impact of oversampling algorithms in the classification of guillain-barré syndrome main subtypes”, Ingenius. Revista de Ciencia y Tecnología, no. 25, pp. 20–31, 2021.

[49] T. K. Ho, “Random decision forests”, in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 1995, vol. 1, pp. 278–282, 1995.

[50] D. R. Cox, “The Regression Analysis of Binary Sequences”, Journal of the Royal Statistical Society: Series B (Methodological), vol. 20, no. 2, pp. 215–232, 1958.

[51] D. J. Hand and K. Yu, “Idiot’s Bayes—not so stupid after all?”, International statistical review, vol. 69, no. 3, pp. 385–398, 2001.

[52] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system”, In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, 2016.

[53] R. Obiedat et al., “Sentiment analysis of customers’ reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution”, IEEE Access, vol. 10, pp. 22260–22273, 2022.

[54] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh, and D. Javed, “Framework for improved sentiment analysis via random minority oversampling for user tweet review classification”, Electronics (Basel), vol. 11, no. 19, p. 3058, 2022.