سال ۱۴۰۰، شماره ۱، صفحات ۴

پردازش زبان طبیعی

Course No. 40677

تشخیص اخبار تاثیرگذار

حسن حمیدی و حامد همتیان و آرش لگزیان [۱ و *]، معین سلیمی [۲]، احسان‌الدین عسگری [۳]

[۱] دانشجوی کارشناسی ارشد هوش مصنوعی دانشگاه صنعتی شریف، [۲] دانشجوی دکتری هوش مصنوعی دانشگاه صنعتی شریف

[۳] استادیار دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف

*مسئول مکاتبات: lagzian@ce.sharif.edu

◄ واژگان کلیدی

تشخیص اخبار تاثیرگذار
پردازش زبان طبیعی
یادگیری ماشین
یادگیری عمیق

◄ تاریخچه مقاله

تاریخ پایان نگارش: ۱۴۰۰/۱۱/۲۹

◄ چکیده

امروزه با وجود حجم زیاد اخبار منتشر شده در خبرگذاری‌ها و رسانه‌های مختلف، بررسی همه آن‌ها برای اکثر افراد جامعه امری سخت و حتی غیرممکن می‌باشد. به همین جهت مشخص کردن اخباری که برای تعداد قابل توجهی از افراد جامعه مهم باشد یک وظیفه[a] مهم درنظرگرفته می‌شود که از آن با نام تشخیص اخبار تاثیرگذار[b] یاد می‌شود. تشخیص اخبار تاثیرگذار توسط انسان فرآیندی زمان‌بر[c] و هزینه‌بر است. از همین‌رو در این پروژه به انجام وظیفه تشخیص اخبار تاثیرگذار توسط مدل‌های هوشمند می‌پردازیم. امروزه مدل‌های هوشمند مبتنی بر یادگیری ماشین[d] و یادگیری ژرف[e] در حوزه‌های مختلف پردازش زبان طبیعی[f] به موفقیت‌های چشمگیر دست یافته‌اند و در این پروژه به بررسی رویکردهای مختلف و ارائه مدل‌های هوشمند با دقت قابل قبول برای انجام این وظیفه می‌پردازیم.

[a] task [b] important news detection [c] time consuming [d] machine learning [e] deep learning [f] natural language processing

۱ مقدمه

اخبار منتشر شده در رسانه‌های مختلف دارای دسته‌بندی‌های مختلفی مانند: اجتماعی، سیاسی، بین‌المللی، ورزشی، حوادث، اقتصادی و ... می‌باشد. در هریک از این دسته‌بندی‌ها تعداد زیادی از اخبار مختلف منتشر می‌شوند که همه این اخبار برای مخاطبین این دسته‌ها مهم نیستند و از همین رو افراد علاقمند هستند تا ابتدا اخبار مهم و تاثرگذار را مطالعه کنند و سپس در صورت داشتن زمان کافی به مطالعه اخبار غیرمهم بپردازند. [۱]، [۲]، [۳]، [۴]، [۵]

۲ تعریف مفاهیم و مسئله

در این بخش ابتدا مفاهیم مورد نیاز تعریف شده سپس مسئله شرح داده شده است.

۱.۲ مفاهیم مورد نیاز

۱. **نمونه خبر تاثیرگذار:** *نمونه‌ای که از توزیع خبرهای تاثیرگذار یا مهم یا داخل دسته برداشته شده است.*

۲. **نمونه خبر غیر تاثیرگذار:** *نمونه‌ای که متعلق به توزیع داده‌های تاثیرگذار نیست.*

۳. **تشخیص خبر تاثیرگذار:** *بطور کلی تمایز قائل شدن بین نمونه خبر تاثیرگذار و غیر تاثیرگذار است.*

۴. **نمونه صحیح ـ مثبت[۱] یا TP:** *نمونه تاثیرگذاری که به درستی تشخیص داده شده است.*

۵. **نمونه صحیح ـ منفی[۲] یا TN:** *نمونه غیرتاثیرگذاری که به درستی تشخیص داده شده است.*

۶. **نمونه اشتباه ـ مثبت[۳] یا FP:** *نمونه غیرتاثیرگذاری که به اشتباه تاثیرگذار تشخیص داده شده است.*

۷. **نمونه اشتباه ـ منفی[۴] یا FN:** *نمونه تاثیرگذاری که به اشتباه غیرتاثیرگذار تشخیص داده شده است.*

۲.۲ بیان مسئله

در این پژوهش قصد داریم تا با تهیه یک مجموعه دادگان برچسب‌دار از اخبار خبرگذاری‌های فارسی مختلف و سپس آموزش مدل‌های هوشمند به حل این چالش بپردازیم.

مجموعه دادگان: جهت انجام وظیفه تشخیص اخبار تاثیرگذار فارسی تا کنون مجموعه دادگانی تهیه و ارائه نشده است و در این پژوهش برای اولین بار یک مجموعه دادگان حاوی چهارهزار نمونه اخبار از خبرگذاری‌های مختلف تهیه و توسط چهار نفر اعضای انجام دهنده این پروژه به صورت دو به دو برچسب‌گذاری شده است و سپس یک نفر از گروه دیگر به‌عنوان نفر سوم بر روی نمونه‌هایی که در هرگروه اختلاف داشتند نظر داده است به این صورت سعی شده است تا حد امکان از بایاس شدن برچسب‌ها بر روی نظر شخصی افراد جلوگیری شود و تا حد امکان به نتایج به نتایج دنیای واقعی نزدیک شود.

[1] true positive [2] true negative [3] false positive [4] false negative

معیارهای ارزیابی: معیارهای ارزیابی در این حوزه موارد زیر می‌باشند:

$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$
$$F1 - Score = \frac{2*Precision*Recall}{Precision*Recall}$$

همچنین از معیار AUROC نیز جهت ارزیابی استفاده شده است. این معیار بر اساس منحنی ROC[5] و ناحیه زیر نمودار AUC[6] تعیین می‌گردد. محور افقیFPR[7] و محور عمودیTPR[8] می‌باشد و با تغییر یک آستانه[9] منحنیROC رسم می‌شود و داشتن ناحیه زیر نمودار بزرگ‌تر به معنی عملکرد بهتر است.

$$TPR = \frac{TP}{TP+FN}$$
$$FPR = \frac{FP}{TN+FP}$$

۳ روش‌های معرفی شده

۱.۳ رویکردهای مبتنی بر یادگیری ماشین کلاسیک

در این بخش با استفاده از پنج مدل یادگیری ماشین کلاسیک به مسئله تشخیص اخبار تاثیرگذار پرداخته شده است. در این مدل‌ها ترکیب‌های مختلفی برای ورودی در نظر گرفته شده است اما بهترین نتایج با در نظر گرفتن ستون داده و استفاده از جستجوی شبکه‌ای[10] و یافتن بهترین ابرپارامترها[11] برای داده ورودی برای هر یک از مدل‌ها با محوریت معیار ارزیابی f1-macro در جدول ۱.۳ قابل مشاهده هستند. برای آموزش و اعتبارسنجی این مدل‌ها از تکنیک k-fold cross validation نیز استفاده شده است.

Metrics	Before Tuning		After Tuning	
	TF-IDF	BoW	TF-IDF	BoW
Logistic Regression	۴۷/۷ %	۶۶/۲ %	۶۴/۹ %	۶۵/۳ %
Support Vector Machine	۴۸/۶ %	۴۶/۰ %	۶۴/۹ %	۶۴/۳ %
Naive Bayes	۴۶/۰ %	۵۰/۳ %	۴۶/۰ %	۵۰/۳ %
Decision Tree	۵۵/۹ %	۵۴/۳ %	۵۶/۸ %	۵۵/۳ %
Random Forest	۴۵/۹ %	۴۶/۸ %	۵۵/۳ %	۵۴/۲ %

۲.۳ رویکردهای مبتنی بر شبکه‌های عصبی عمیق

در این رویکرد از مدل BiLSTM Attention Based استفاده شده است. در این مدل تعداد کلمات مجموعه داده ده هزار کلمه فرض شده و بر اساس ده هزار کلمه پرتکرار، داده تبدیل به دنباله‌های ورودی مدل می‌شوند و بقیه کلمه‌هایی که در آن ده هزار کلمه پرتکرار نیست را توکن [12] خارج از محدوده در نظر گرفته می‌شوند. بعد از تبدیل به دنباله، داده‌ها به سه بخش آموزش، ارزیابی و تست با نسبت ۶.۰ و۲.۰ و ۲.۰ تقسیم می‌شوند. شبکه‌ای که آموزش داده شده دارای هسته BiLSTM است که دارای لایه نهان به طول ۱۲۸ و لایه سلولی به طول ۱۲۸ است تمام ورودی‌ها هم به طول ۵۱۲ نرمال می‌شوند، بر روی این شبکه از دو لایه توجه استفاده شده که یک توجه میزان توجه بردار نهان پیش‌رو [13] را با خروجی ها حساب می‌کند و توجه دوم میزان توجه بردار نهان عقب‌رو [14] را با خروجی محاسبه می‌کند. شبکه‌ای که برای توجه به کار می‌رود یک لایه خطی با تعداد نرون ۵۰ است. در نهایت بردارهای حاصل از این توجه ها را به هم متصل می‌شوند و به یک لایه تماما متصل داده می‌شوند این لایه تماما متصل دارای ۵۱۲ نرون است و از relu برای فعال‌سازی و در زمان آموزش از تکنیک

drop-out با احتمال ۰/۸ نیز استفاده شده است. از آنجایی که تعداد داده‌ها کم است این میزان از حذف نرون احتمال overfit را کاهش می‌دهد و جدای از این مورد به نوعی تکنیک ensemble را نیز در این لایه شبیه‌سازی می‌کند که برای داده‌های نامتوازن مثل داده‌های این مجموعه دادگان می‌تواند مناسب باشد، بعد از آن دو لایه با تعداد ۵۰ نرون و ۲ نرون استفاده می‌شود. لایه فعال‌ساز این دو لایه relu و softmax انتخاب شده‌اند. در فرایند آموزش این شبکه از هزینه cross entropy استفاده شده چون داده‌ها نامتوازن است میزان هزینه به صورت وزن‌دار حساب می شود، برای داده‌های غیرمهم وزن ۰/۵ و برای داده‌های مهم وزن ۱ گذاشته شده این مقادیر طبق تجربه بدست آمده است. آموزش این شبکه سریع است و تنها به سه دور [15] برای آموزش نیاز دارد که هر چرخه در حدود ۱۰ ثانیه زمان نیاز دارد. نتایج پیاده سازی این مدل در قسمت نتایج قابل مشاهده است.

۳.۳ رویکردهای مبتنی بر شبکه‌های ترنسفورمری

امروزه مدل‌های ترنسفورمری [16] در اکثر زمینه‌های مرتبط به یادگیری ماشین و پردازش زبان طبیعی در جایگاه اول قرار دارند و استفاده از آنها برای وظیفه‌های مختلف بسیار مورد توجه است. در این پژوهش با استفاده از مدل‌های ترنسفورمری روبرتا[17] و برت[18] به حل مسئله تشخیص اخبار تاثیرگذار نیز پرداخته شده است. ورودی این مدل‌ها مانند سایر رویکردهای قبلی، داده‌های پیش پردازش شده هستند که با ترکیب‌های مختلفی مثل در نظر گرفتن متن، عنوان، متن و عنوان، متن و دسته‌خبر، متن و نام خبرگذاری و ... کارایی مدل مورد ارزیابی قرار گرفته است و نتایج حاصل از استفاده از این مدل‌ها در بخش نتایج تحلیل می‌شوند.

برای بهبود عدم توازن داده تاثیرگذار و غیرتاثیرگذار نیز در شبکه روبرتا از دو رویکرد Oversampling

اخبار تاثیرگذار و KMeans Sampling اخبار غیر تاثیرگذار استفاده شد. در Oversampling داده‌های تاثیرگذار تا نصف داده‌های غیرتاثیرگذار افزایش می‌یابند و شبکه روبرتا با داده‌های جدید آموزش می‌بینند. در KMeans Sampling داده غیرتاثیرگذار با الگوریتم KMeans به خوشه‌هایی به تعداد داده‌های تاثیرگذار تقسیم می‌شوند سپس در هر batch تعدادی داده تاثیرگذار و به همان اندازه تعدادی خوشه‌ی غیر تاثیرگذارانتخاب شده و از هر خوشه بصورت تصادفی یک داده غیر تاثیرگذار در batch قرار می‌گیرد. بدین‌صورت تعداد داده‌های مهم و غیرمهم در هر batch برابر می‌شود.

۴ نتایج

[5] receiver operating characteristic [6] area under curve [7] false positive rate [8] true positive rate [9] threshold [10] grid search [11] hyperparameters [12] token
[13] forward [14] backward [15] epoch [16] transormers models [17] Roberta [18] Bert

۱.۴ نتایج مدل‌ها

Model	F1-Macro
Naive Bayes	۵۰.۳ %
Random Forest	۵۵.۳ %
Drecision Tree	۵۶.۸ %
Logistic Regression	۶۴.۹ %
Support Vector Machine	۶۴.۹ %
Bilstm	۶۲.۴ %
Bert	۶۲.۴ %
Roberta	۶۴.۳ %
Roberta + KMeans Sampling	۶۰.۰ %
Roberta + Oversampling	۶۴.۸ %
Roberta + Keywords	۶۲.۶ %
Roberta + Category	**۶۶.۸ %**
Logistic Regression + Bilstm + Roberta	**۶۷.۵ %**

همانطور که ملاحظه می‌کنید استفاده از متن به علاوه دسته‌بندی به عنوان ورودی شبکه روبرتا توانسته بهترین دقت را روی دادگان تست بدست آوزد که این می‌تواند نشان‌دهنده‌ی این موضوع باشد که مدل روبرتا احتمالا نمی‌تواند دسته‌بندی خبر را بدرستی از روی متن تشخیص دهد اما این دسته‌بندی خبر تاثیر بسزایی در اهمیت یا عدم اهمیت اخبار می‌گذارد.
از طرف دیگر در اینجا و شبکه رورتا KMeans Sampling از اخبار غیرمهم نتوانسته تاثیری خوبی در بهبود مدل داسته باشد اما OverSampling توانسته نسبت به حالت عادی روبرتا دقت را مقدار کمی بهبود دهد.
همچنین مدل ensemble توانسته بهترین نتیجه را درکل بدست آورد.

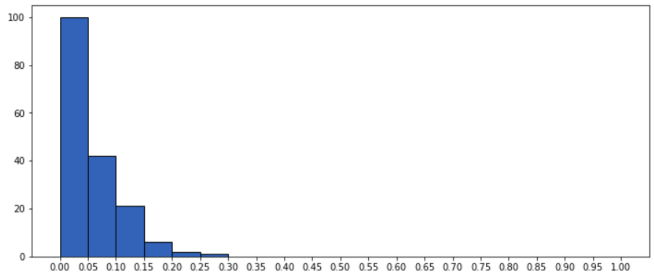
۲.۴ نتایج آزمایش‌ها

ما ابتدا تاثیر بایاس بودن مدل‌ها به نام خبرگذاری را آزمایش می‌کنیم. برای آزمایش این موضوع ما یکبار نام خبرگذاری را از متن خبر حذف کرده و آزمایش می‌کنیم که دقت نسبت به حالت پایه به چه صورت تغییر می‌کند.

	متن	متن بدون نام خبرگذاری
Logistic Regression	۶۴.۹ %	۶۴.۶ %
Support Vector Machine	۶۴.۹ %	۶۴.۷ %
Bilstm	۶۲.۴ %	۶۲.۳ %
Bert	۶۴.۳ %	۶۳.۹ %

همانطور که مشاهده می‌شود پس از حذف نام خبرگذاری از متن و آموزش دقت نهایی رو داده تست برای تمامی مدل‌ها پایین‌تر می‌آید و نتیجه می‌شود که در داده اصلی وجود نام خبرگذاری باعث بایاس شدن مدل نشده است.

مورد دیگر که مورد آزمایش قرار گرفته است این است که توزیع خطا داده‌های دادگان تست که به مدل آن‌ها را اشتباه پیش‌بینی می‌کند به چه صورت است. این موضوع از این جهت مهم است چرا که در هنگام برچسب‌زنی اخبار، بسیاری از اخبار در مرز مهم بودن و غیرمهم بودن قرار می‌گرفتند و تصمیم‌گیری برای آن‌ها پیچیده بود.
شکل پایین توزیع مقدار خطای نمونه‌های اشتباه در مدل Logistic Regression است. در فایل scoring.py نتایج برای دو مدل دیگر نیز مشخص است.



همانگونه که مشخص است بسیاری از اخباری از دادگان تست که به اشتباه برچسب خورده‌اند دارای خطای پایین و بسیاز نزدیک به مرزهای دسته‌بند هستند و ما احتمالا از این آزمایش می‌توانیم این نتیجه‌گیری را انجام دهیم که دقت این مسئله را به دلیل این ابهام نتوان مقدار زیادی افزایش داد.

موضوع دیگری که ما روی آن آزمایش انجام دادیم این است که ارتباط دسته اخبار با مهم بودن اخبار به چه صورت است. برای این کار ما این ارتباط را روی سه مدل Logistic Regression و Bilstm و Roberta آزمایش کردیم. برای بدست آوردن این ارتباط از Mutual Information بین دسته اخبار ورودی و برچسب‌های پیش‌بینی شده مدل استفاده شده است. در شکل زیر برای مدل Logistic Regression این ارتباط را مشاهده می‌کنید.

```
Logistic Regression Category Mutual Information
0.029697166109571738 <-- اجتماعی
0.026101939468282875 <-- اقتصادی
0.023554850151957796 <-- بین الملل
0.018625982350700676 <-- حوادث
0.010467847370794425 <-- سیاسی
0.01435428777387604 <-- علمی و پزشکی
0.03302084623623136 <-- فرهنگ و هنر
0.01595202012622865 <-- فناوری و ارتباطات
0.009801944831982379 <-- مذهبی
0.022205372752022239 <-- ورزشی
```

استخراج میزان اهمیت توکن‌ها: برای تشخیص اهمیت هر توکن برای اخبار تاثیرگذار و غیرتاثیرگذار از mutual information توکن‌های ورودی شبکه روبرتا و برچسب پیش‌بینی شده شبکه در داده‌های آموزش استفاده می‌کنیم. در فایل Roberta News FInal برای هر کلاس ۴۰ توکن با بیشترین mutual information با آن کلاس استخراج شده و نمایش داده شده اند.

۳.۴ پیشنهادات جهت کارهای آتی:

۱. با توجه به پیچیدگی وظیفه تشخیص ناهنجاری استفاده از داده‌های بیشتر با تعداد بیشتری از رای دهندگان برای برچسب‌گذاری داده‌ها می‌تواند تاثیر مثبتی بر نتایج مدل‌ها و عمومیت آن‌ها داشته باشد.

۲. استفاده از رویکردهای نیمه‌نظارتی [۱۹] می‌تواند جهت افزایش مجموعه دادگان بکار گرفته شود.

[19] semi supervised

۵ نتیجه‌گیری

در این پروژه برای اولین بار مجموعه دادگانی به زبان فارسی برای تشخیص اخبار تاثیرگذار ایجاد شد. مدل‌های مختلف یادگیری کلاسیک و مدرن روی مجموعه داده آزمایش شد و نتایج گزارش شد.

آزمایش‌هایی که انجام شد حاکی از این موضوع است که این مسیله دارای ابهام زیادی است که نمودار توزیع خطای دیتای اشتباه برچسب زده شده به این موضوع دلالت دارد. بطور کلی شبکه روبرتا با اعمال تغییرات از جمله اضافه کردن دسته خبر به ورودی توانست بهتر از تمامی مدل‌های دیگر از جمله خودش عمل کند و این احتمالا به این موضوع دلالت دارد که شبکه روبرتا که روی متن اخبار آموزش دیده نمی‌تواند به خوبی دسته خبر را تعیین کند اما این دسته در افزایش دقت مهم است به طوری که زمانی که دسته خبر را به علاوه متن به ورودی می‌دهیم دقت شبکه 4.2 % افزایش می‌یابد.

با ذکر این موضوعات به نظر می‌رسد که برای کارهای آینده با آزمایش بیشتری روی دسته اخبار صورت گیرد. همچنین باتوجه به اینکه این مساله از کمبود داده‌های مهم رنج می‌برد احتمالا یکی از رویکردهای آینده باید حل مشکل عدم توازن داده خبر مهم با داده خبر غیرمهم از طریق یا تولید داده تاثیرگذار بیشتر و یا استفاده از روش‌هایی مانند یادگیری نیمه‌نظارتی باشد.

مراجع

[1] Mostafavi, Sareh, Pahlevanzadeh, Bahareh, and Falahati Qadimi Fumani, Mohammad Reza. Classification of persian news articles using machine learning techniques. *Computer and Knowledge Engineering*, 3(1):73–81, 2020.

[2] Davari, Nafiseh, Mahdian, Mahya, Akhavanpour, Alireza, and Daneshpour, Negin. Persian document classification using deep learning methods. in *2020 28th Iranian Conference on Electrical Engineering (ICEE)*, pp. 1–5. IEEE, 2020.

[3] Rezaeian, Naeim and Novikova, Galina. Persian text classification using naive bayes algorithms and support vector machine algorithm. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 8(1):178–188, 2020.

[4] Ghasemi, Saeideh and Jadidinejad, Amir H. Persian text classification via character-level convolutional neural networks. in *2018 8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN)*, pp. 1–6. IEEE, 2018.

[5] Varasteh, Mohammadreza and Kazemi, Arefeh. Using parsbert on augmented data for persian news classification. in *2021 7th International Conference on Web Research (ICWR)*, pp. 78–81. IEEE, 2021.