تشخیص موضع کاربران در مورد شایعه در شبکه های اجتماعی

دانشجو: سارینا جامی الاحمدی

استاد راهنما: دکتر شریعت پناهی

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

مقدمه

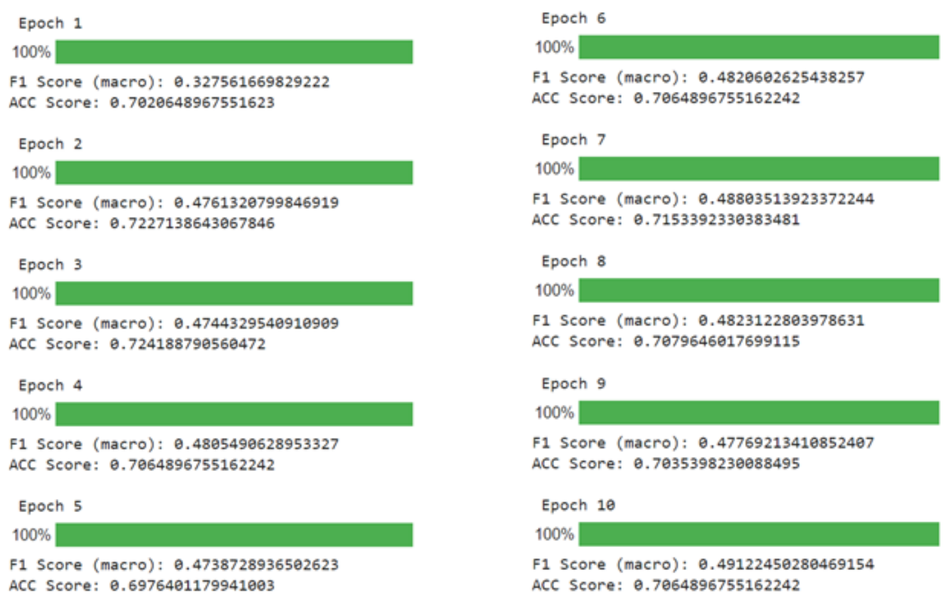
امروزه به دلیل پیشرفت تکنولوژی و در دسترس بودن اینترنت، رسانه های اجتماعی به عنوان یکی از اصلی ترین بستر های مجازی برای برقرار کردن ارتباطات، به اشتراک گذاشتن اطلاعات و گسترش آن هستند. با وجود اینکه شبکه های مجازی باعث ایجاد رفاه بیشتری شده اند ولی مشکلاتی را نیز به همراه آورده اند. یکی از این مشکلات، مسئله پخش شدن شایعه در شبکه های اجتماعی و آسیب رساندن آن به افراد و جامعه است. به منظور صحت سنجی شایعه، پژوهش های مختلفی انجام شده است که این تحلیل ها نشان داده استفاده از نظرات کاربرانی که در مورد صحت شایعه به گفتگو میپردازند میتواند کمک شایانی در تشخیص آن بکند. کاربران میتوانند با تایید کردن و رد کردن شایعه و ارائه شواهدی برای آن از اعتبار خبر اطلاع دهند. پژوهشگران با استفاده از روش های یادگیری ماشین به طبقه بندی موضع کاربران جهت بهبود بخشیدن به یک سیستم تشخیص شایعه میپردازند. ما نیز در این پژوهش قصد داریم ابتدا با مطالعه دقیق تر این حوزه و اطلاع از چالش ها و پیچیدگی های آن به پیاده سازی یک مدل شبکه عصبی عمیق به منظور طبقه بندی موضع کاربران در مورد شایعه بپردازیم.

مدل پیشنهادی

با توجه به مطالعات انجام شده، موضع کاربران در مورد شایعه به چهار دسته تقسیم میشود که شامل تایید کردن، رد کردن، سوال پرسیدن و نظر دادن است. به منظور طبقه بندی موضع کاربران، از مدل pre-trained BERT و مجموعه داده RumourEval-2017 که اولین مجموعه داده مرتب شده برای تشخیص شایعه بوده است استفاده میکنیم.

مدل BERT یک مدل پردازش زبان طبیعی است که توسط تیم هوش مصنوعی شرکت گوگل در سال ۲۰۱۸ توسعه یافته است. این مدل اولین مدلی است که یادگیری مفهوم را به صورت دو طرفه انجام میدهد که باعث میشود فهم عمیق تری از متن بدست آورد. لایه های اولیه مدل BERT توسط مجموعه داده های بزرگی یاددهی شده است و بدین صورت عمل میکند که یاددهی چند لایه آخر بر روی مجموعه داده مورد نظر ما صورت میگیرد و این مدل وزن های جدیدی مرتبط با داده های ما یاد میگیرد.

برای یاددهی این مدل ما از متن توییت ها و شناسه آن ها که مشخص کننده این است توییتی به توییت دیگر پاسخ داده است یا خیر استفاده کردیم. سپس، به پیش پردازش متون جهت آماده سازی آن برای ورودی مدل پرداختیم. این ورودی را به الگوریتم BERT for sequence classification که در لایه آخر آن از یک مدل طبقه بندی خطی استفاده شده است میدهیم. به منظور کاهش هزینه محاسباتی و عملکرد بهتر از الگوریتم Mini batch gradient descent بهره میبریم. سپس برای یاددهی مدل تعداد یک تا ده epoch اجرا میکنیم و در هر مرحله نتایج accuracy و f1-score را بر روی داده تست بدست میاوریم. همانطور که در شکل زیر مشاهده میکنیم تفاوت چشم گیری در f1-score بین epoch اول و دوم وجود دارد که میتواند دلیل آن به خوبی مدل نشدن مجموعه داده باشد. از epoch دوم تغییر محسوسی مشاهده نمیشود و تا epoch دهم داده ها overfit نمیشوند زیرا که کاهش چشم گیری مشاهده نمیکنیم.



نتایج

در ارزیابی نتایج مدل های طرح شده برای دسته بندی موضع کاربر معیار f1-score اعتبار بیشتری نسبت به accuracy دارد زیرا که این مسئله باینری نیست و همچنین داری عدم توازن زیادی در مجموعه داده است.

همانطور که در جدول زیر مشاهده میکنیم مدل ما توانسته به نتایج قابل قیاسی در معیار f1-score با مقاله های ۱، ۲ و ۴ برسد. در این مسئله کلاس کم اهمیت تر (نظر دادن) که بیشترین نمونه را دارد معمولا با دقت خوبی توسط مدل های زیادی پیش بینی شده است و accuracy را افزایش میدهد این در حالی است که مدل های مربوطه دقت بسیار کمی در رابطه با کلاس مهم تر (رد کردن) دارند زیرا که تعداد نمونه های آن بسیار اندک است و این باعث کاهش معیار f1-score برای آن ها میشود.

ما در این پژوهش تنها از داده های متن برای دسته بندی موضع کاربران استفاده کردیم و همانطور که مشاهده میکنیم نتیجه خوبی به ما داده است. برای بهبود بخشیدن به عملکرد مدل، در قدم های بعدی پژوهش از ویژگی های دیگر مجموعه داده مانند اطلاعات کاربر نیز بهره خواهیم برد تا دقت مدل را افزایش دهیم.

Reference	Model	Macro F1-score	Accuracy
Pamungkas *et al.* [1]	SVM	47%	79.5%
Veyseh *et al.* [2]	CNN	48.2%	82%
Kumar *et al.* [3]	Capsule Network and Multi-layer Perceptron	55%	77.6%
Yu *et al.* [4]	Hierarchical Transformer with BERT	50.9%	76.3%
Our Model	Pre-trained BERT	49.12%	70.65%

جمع بندی

تشخیص شایعه های نادرست در شبکه های اجتماعی باعث جلوگیری از آسیب دیدن افراد و جامعه میشود. با بکارگیری مدل های یادگیری مطرح شده در شبکه های اجتماعی میتوانیم کاربران را از اخبار نادرست مطلع کنیم و همچنین باعث آشنایی بیشتر آنان با شایعه هایی که نامتعبر هستند شویم. این مسئله چالش ها و محدودیت های متعددی دارد مانند نبود یک سیستم داینامیک کارآمد، کم بودن مجموعه داده های غیر انگلیسی، تعداد بسیار اندک داده های مربوط به کلاس رد کردن که مهمترین کلاس است و نیاز به توسعه مدل های نیمه نظارت شده و نظارت نشده به دلیل زمانبر و پرهزینه بودن برچسب گذاری داده ها. این حوزه در حال حاضر بسیار مورد توجه پژوهشگران قرار گرفته و در تلاش برای طراحی یک سیستم موثر هستند.

مراجع اصلی

1. E. W. Pamungkas, V. Basile, and V. Patti, “Stance classification for rumour analysis in Twitter: Exploiting affective information and conversation structure,” ArXiv, vol. abs/1901.01911, 2019.
2. A. P. B. Veyseh, J. Ebrahimi, D. Dou, and D. Lowd, “A temporal attentional model for rumor stance classification,” in Proc. ACM Conf. on Info. and Knowl. Manage. (CIKM), Nov. 2017, pp. 2335–2338.
3. A. Kumar and M. Upadhyay, “Rumor stance classification using a hybrid of capsule network and multi-layer perceptron,” Turkish J. of Comput. and Math. Educ. (TURCOMAT), vol. 12, no. 13, pp. 4110–4120, 2021.
4. J. Yu, J. Jiang, L. M. S. Khoo, H. L. Chieu, and R. Xia, “Coupled hierarchical transformer for stance-aware rumor verification in social media conversations,” in Proc. Conf. on Empirical Methods in Natural Lang. Process. (EMNLP), Nov. 2020, pp. 1392–1401.