Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال اول، شماره دوم، پاییز و زمستان ۱۴۰۲، صفحات ۱۳۰-۱۱۸

تشخیص مکاتب فلسفی با استفاده از یادگیری عمیق

احمد علی محمدی[۱]، محمدمهدی حسینی [۲*]

[۱]دانشکده مهندسی کامپیوتر، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران

[۲]دانشکده مهندسی کامپیوتر، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران

مقاله پژوهشی

تاریخ دریافت:

۱۴۰۲/۰۴/۰۴

تاریخ پذیرش:

۱۴۰۲/۰۸/۳۰

کلیدواژه‌ها:

مکاتب فلسفی، داده کاوی متن، تشخیص مکاتب، تبدیل متن به بردار عددی، میزان شباهت تفکر نویسنده با مکاتب

نویسنده‌مسئول:

hosseini_mm@yahoo.com

چکیده

در حوزه شناخت مکاتب فلسفی، تفکر هر شخص با توجه به نوع نگرشی که در خصوص مکاتب مختلف می تواند داشته باشد متفاوت است. تشخیص نگرش نویسنده و میزان شباهت آن به هرکدام از مکاتب فلسفی همواره یکی از موضوعات مهم در حوزه علوم انسانی بوده است. در این مقاله، روشی مبتنی بر یادگیری عمیق برای تشخیص مکاتب فلسفی از روی متن پیشنهاد شده است. در روش پیشنهادی، ابتدا متن ها نرمال شده و کلمات اضافی و فاقد معنا حذف می شوند. بعد از مرحله نرمال سازی، متن به جملات و کلمات شکسته شده و سپس با استفاده از کتابخانه فست تکست هر کلمه به بردار عددی تبدیل می شوند، پس از آن با استفاده از شبکه طراحی شده ویژگی های متون استخراج شده و در نهایت سیستم، متون را یاد گرفته و آماده برای استخراج داده ها می باشد و با دادن یک جمله جدید میزان شباهت آن به هر مکتب بیان می شود. بر اساس ارزیابی صورت گرفته، میزان دقت در روش پیشنهادی ۹۴ درصد می باشد.

: 10.22034/ABMIR.2023.20250.1029

E-ISSN: 2821-2037 /© 2023. Published by Yazd University This is an open access article under the CC BY 4.0 License (https://creativecommons.org/licenses/by/4.0/).

۱- مقدمه

فلسفه، واژه‌ای با ریشه یونانی و معرب فیلوسفیا به معنای دوستدار دانش است[۱]. به بیانی دیگر در فلسفه فعالیت فکری و ذهنی در زمینه موضوعات و مسائل بنیادینی از حیطه علوم تجربی خارج بوده و با روش تجربی قابل اثبات یا رد نبوده و تابع قرارداد و اعتبار نیست و با روش نقلی اثبات نمی‌شود. بنیادی‌ترین مسئله در اندیشه نوع بشر، پرسش از هستی و واقعیت است که قدمتی به درازای اندیشیدن و تعقل دارد. به همین مناسبت فیلسوفان مسلمان از ابونصر فارابی و ابن سینا تا فلاسفه متاخر، همگی فلسفه را دانشی دانسته‌اند که در آن از احوال هستی و واقعیت از جهت واقعیت‌مندی یا پنداری بودن آن، مورد بحث قرار می دهند[۳][۲]. در تحقیقات گذشته در زمینه علوم انسانی مکاتب فلسفه، از دیدگاه های مختلف مورد بررسی قرار گرفته است. گرایش های متفاوتی در مکتب های فلسفی از قبیل فلسفه سیاسی، فلسفه اجتماعی، فلسفه اخلاقی و غیره که هرکدام دارای ایدئولوژی منحصر به فرد خود هستند وجود دارد. با ظهور هوش مصنوعی در قرن جدید می توان در بسیاری از زمینه‌ها از این تکنولوژی بهره گرفت. در این راستا، چهار مکتب فلسفی سوسیالیسم، سکولاریسم، لیبرالیسم و مارکسیسم، برای ارائه روش پیشنهادی مورد استفاده قرار گرفت. امروزه استفاده از روشهای مبتنی بر یادگیری عمیق در هنگام کار با متون بسیار بزرگ و حجیم یک رویکرد جذاب و کاربردی است. یادگیری عمیق و تمام جنبه‌های هوش مصنوعی مدرن از داده‌ها استفاده می‌کنند تا تصمیم‌های مشابه هوش انسانی بگیرند. یکی از کاربردهای اصلی یادگیری عمیق، استفاده از آن در حوزه داده کاوی می باشد. امروزه با پیشرفت های صورت گرفته در پردازش اطلاعات و همچنین دسترسی رایگان به متون دیجیتالی، بررسی حجم عظیمی از متون فلسفی به منظور یافتن اطلاعاتی که در این متون هستند به سادگی امکان پذیر می باشد. با استفاده از یادگیری عمیق و تکنیک های داده کاوی میتوان حجم عظیمی از داده و اطلاعات را به سیستم آموزش داده و اطلاعاتی سودمند بر اساس نیاز استخراج و بهره بردای های لازم را انجام داد. آنچه در این مقاله بررسی می شود یادگیری متون مکاتب فلسفی و سپس میزان شباهت متون جدید به هر کدام از مکاتب بر اساس آنچه یاد گرفته شده است می باشد. از جمله نتایج حاصل از این مقاله به، بهره وری کامل تر از منابع موجود در مکاتب فلسفی، استفاده از منابع بیشتر در مدت زمان کمتر، استفاده سریع تر و آسان تر از منابع، مقایسه و شناخت شباهت متون با یکدیگر، تشخیص متن نویسنده و شناخت شباهت متن به متون موجود در پایگاه داده میتوان اشاره کرد. در این مقاله در بخش دوم مروری بر کارهای گذشته صورت گرفته است. در بخش سوم به بیان روش پیشنهادی خواهیم پرداخت. در بخش چهارم ارزیابی از روش پیشنهادی خواهیم داشت و در فصل پنجم به نتیجه گیری و کارهایی که در آینده قابل انجام خواهد بود اشاره شده است.

۲- مروری بر کار های گذشته

در پژوهشی باتوجه به اهمیت تکنیک‌های موثر و کارآمد پردازش زبان، به مروری بر روندها و تکنیک های پردازش زبان طبیعی با استفاده از NLTK پرداخته شده است[٤]. در پژوهشی دیگر به محاسبه خودکار معیارهای غنی واژگانی برای یادگیری L2 با مقایسه ابزارهای LCA، NLTK و SpaCy پرداخته شده است[٥]. همچنین برای طبقه بندی موضوعی اسناد با استفاده از مجموعه ای از مقالات ویکی پدیا روشی مبتنی بر یک شبکه LSTM با اسنادی که به صورت دنباله ای از کلمات کدگذاری شده و با بردارهای ورد تو وک[1] ایجاد گردیده ارائه نموده است. ارزیابی و آزمایشات آنها نشان داد که رویکرد ارائه شده نسبت به یک رویکرد استاندارد و کیفی کلمه با اسناد که از TF-IDF استفاده نموده باشد بهتر عمل کرده است[٦]. در پژوهشی دو تکنیک CBOW و اسکیپ گرام[2] برای تولید مدل‌های جاسازی کلمات، دقیق و کارآمد شناخته شده و در این مورد با بهره‌گیری این دو تکنیک که توسط Mikolov [٧] پیشنهاد شده، یک نسخه کاملاً بهینه شده از این مدلها را با استفاده از Py-Torch پیاده سازی کردند و در تجزیه و تحلیل احساسات در موضوع اسباب بازی بکار گرفتند[٨].

[1] Word2vec

[2] Skip-gram

در پژوهشی دیگر به مقایسه دو، مدل CBOW و مدل اسکیپ گرام پرداخته و نتیجه گرفته، مدل اسکیپ گرام دارای دقت بالایی در نمایش توزیع کلمات است، اما یادگیری آن زمان زیادی طول می‌کشد. از سوی دیگر، مدل CBOW در نمایش توزیع کلمات سریع تر اما از دقت کمتری نسبت به مدل اسکیپ گرام برخوردار است. در این مطالعه، روشی را برای ترکیب مدل CBOW و مدل اسکیپ گرام به شیوه ای پیشنهاد داده که ماتریس وزن مدل CBOW را با مدل اسکیپ گرام به اشتراک می‌گذارد تا هم سرعت یادگیری و هم دقت نمایش توزیع را بدست بیاورد[۹]. در این پژوهش ایده ای، برای حل مشکل خوشه بندی پیشنهاد شده است. که در مرحله اول از جاساز کلمات استفاده شده، که ویژگی معنایی کلمات را به همراه اطلاعات دستوری و نحوی به ارمغان می‌آورد و در مرحله بعد، اسناد به عنوان مجموع وزنی TF-IDF بردارهای تعبیه شده مربوط به هر کلمه نشان داده می شوند. وزن بیشتر TF-IDF به این معنی است که کلمات در اسنادی که در آنها آمده است رابطه قوی تری دارند. در نهایت، با استفاده از خوشه بندی سلسله مراتبی، تمام اسناد نوشته شده توسط هر نویسنده را با یکدیگر گروه بندی می کند[۱۰].

برای دسته بندی متون، هر متن با توجه به عنوان و میزان شباهتی که با داده های هم نوع دارند دسته بندی می شوند. در این روش الگوریتم‌های مختلفی برای این کار وجود دارد که با مرور جامع بیش از ۱۵۰ مدل مبتنی بر یادگیری عمیق، روش هایی برای طبقه بندی متن ایجاد شده را که در سال های اخیر ارائه داده اند و در مورد مشارکت های فنی، شباهت ها و نقاط قوت آنها بحث کرده اند. همچنین خلاصه ای بیش از ۴۰ مجموعه داده محبوب را ارائه نموده اند[۱۱]. در پژوهشی به جهت جلوگیری از سخنان توهین آمیز در قسمت نظرات رسانه های اجتماعی، دسته ای از کلماتی را که در مکالمات بین جوانان رایج بوده را با استفاده از کتابخانه NLP پیش پردازش کرده و این مجموعه داده را با استفاده از جاساز (برت[1] و فست تکست) و الگوریتم‌های یادگیری عمیق (CNN, MLP, LSTM) طبقه بندی کرده است[۱۲]. همچنین پس از بیان این موضوع که طبقه‌بندی متن، بر اساس الگوریتم‌های یادگیری ماشین مرسوم دارای مشکلاتی مانند مشکل ابعاد و کارایی پایین هستند، یک مدل طبقه بندی متن بر اساس فست تکست را پیشنهاد می‌کند، که در این مدل اطلاعات مهم موجود در متن را از طریق مهندسی ویژگی بررسی کرده است. این بررسی از پایتون برای طبقه‌بندی مجموعه داده متنی (تشخیص محتوای عاطفی نظر کاربران)در پلتفرم بایدو دیانشی است[۱۳].

۳- روش پیشنهادی

در روش پیشنهادی ابتدا داده‌ها با استفاده از کتابخانه هضم[2] پیش پردازش می‌شوند، سپس هر متن با توجه به عنوانی که دارد در مجموعه داده مخصوص خود قرار می گیرد و بعد از آن داده ها به دو دسته آموزش و آزمون تقسیم می شوند، در مرحله بعد آنها را تبدیل به بردار های عددی کرده و پس از ان اقدام به استخراج ویژگیهای مورد نیاز از بردارهای عددی ساخته شده بر مبنای مدل می گردد. پس از یادگیری، سیستم ارزیابی شده و نتیجه بررسی می شود. شکل (۱) فلوچارت روش پیشنهادی را نمایش می دهد.

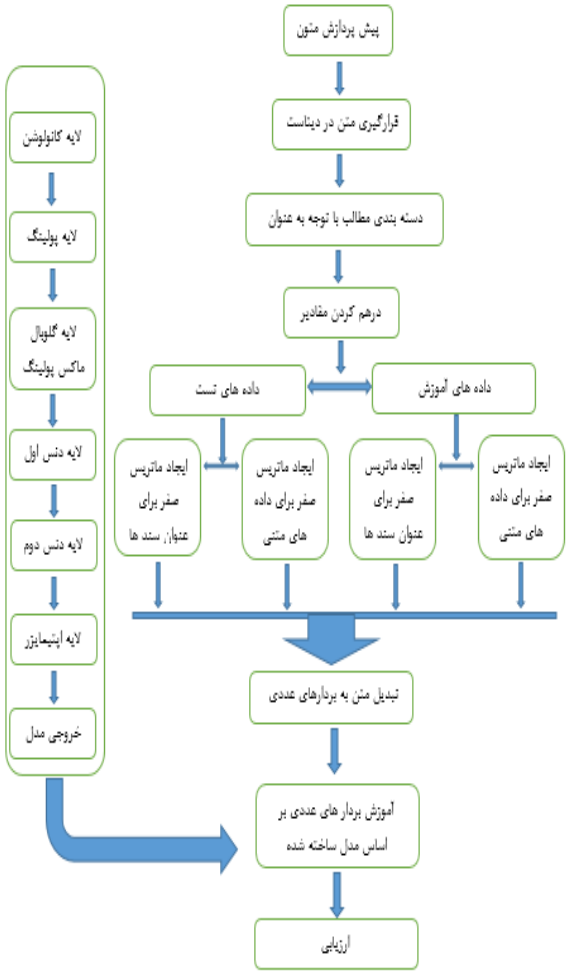
۳-۱- پایگاه داده

با توجه به جدید بودن موضوع در روش پیشنهادی، پایگاه داده استانداری برای این مهم وجود نداشت. بنابراین نیاز به جمع‌آوری متن‌های مرتبط با این موضوع بود. با بررسی‌های صورت گرفته، متون پایگاه داده از پژوهشگاه علوم انسانی و مطالعات فرهنگی جمع‌آوری شد. در مجموعه داده جمع‌آوری شده چهار مکتب مورد بررسی قرار گرفت. جدول ۱ تعداد کلمات مربوط به هر مکتب را نشان می‌دهد.

جدول(۱): تعداد کلمات هر مکتب

نام مکتب	تعداد کلمه
سکولاریسم	۳۰۷۲۰
سوسیالیسم	۳۳۷۱٤
مارکسیسم	۱۵۳۵٤
لیبرالیسم	۲۸٦۹۲

[1] Bert
[2] Hazm



شکل(۱): فلوچارت مراحل پیشنهادی

۳-۲- پیش پردازش متون

متن اولیه را باید به قالبی تبدیل کرد که کلمات دارای معنا استخراج شده، سپس کلمات با معنا برای آموزش به سیستم استفاده گردد. برای نرمال سازی متن، حذف ایست واژه و غیره از کتابخانه ای به نام هضم که بر گرفته از کتابخانه قدرتمند NLTK[1] می‌باشد استفاده شده است.

۳-۲-۱- نرمالسازی متن

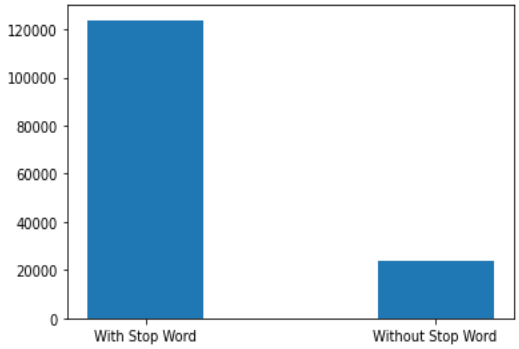
در واقع قبل از پردازش متون جهت استانداردسازی حروف و فاصله‌ها بایستی پیش‌پردازش‌هایی روی آنها انجام شود که در این مرحله، تمیز و مرتب کردن متن و یکسان‌سازی کاراکترها با جایگزین کردن کاراکترهای استاندارد در متن ورودی صورت می پذیرد.

۳-۲-۲- تقطیع جملات و واژه‌ها

این مرحله، تشخیص و استخراج کلمات از متن ورودی است. به عبارت دیگر جداسازی جملات ساده و مرکب (غیرتودرتو) و واژه‌ها و عبارات خاص (توکن‌ها) در این مرحله صورت می پذیرد.

۳-۲-۳- حذف کلمات توقف

حذف کلمات توقف، حذف علائم، اعداد، کلمات عمومی و بدون ارزش معنایی (از قبیل: از، در، با، به، است، پس، و غیره) می‌بایستی انجام پذیرد. برای انجام این کار در ابتدای کار متن را فراخوانی و نرمال سازی کرده و فاصله ها، نیم فاصله های اضافی، کلمات ایست واژه را از جملات حذف کرده (کلماتی که دارای محتوای خاصی نیستند و مفهومی را انتقال نمی دهند) و پس از آن سایر علامت ها و اعداد را از متون حذف می کنیم. شکل ۲ مقایسه تعداد کلمات قبل از مرتب سازی متون و بعد آن را نشان می دهد. همانطور که در شکل مشاهده می شود، تعداد کلمات پس از این مرحله کاهش چشمگیری را نشان می دهد که این کاهش منجر به کاهش بار پردازشی سیستم خواهد شد.



شکل(۲): تعداد کلمات، قبل و بعد مرتب سازی

۳-۳- دسته بندی مطالب و قرارگیری متن در مجموعه داده

در این مقاله از یک روش دسته بندی چهار کلاسه استفاده شده است. به این معنا که میزان احتمال یک متن ورودی به یکی از مکاتب سکولاریسم، سوسیالیسم، لیبرالیسم و مارکسیسم مشخص

[1] Natural language toolkit

می شود. پس از قرار گیری داده ها در پایگاه داده، داده ها به دو دسته، داده هایی برای یادگیری و داده هایی برای آزمون دسته بندی می شود. بدیهی است در صورت استاندارد نبودن داده های آزمون و آموزش نتیجه ارزیابی به خوبی قابل استناد نباشد. در این مقاله ۸۰ درصد داده ها موجود برای اموزش و بقیه برای آزمون در نظر گرفته شده است. به این ترتیب می‌توان میزان دقت الگوریتم را به دست آورد. برای آنکه داده آزمون، ترکیبی از داده های مختلف از مکاتب مختلف فلسفی باشد، مقادیر را با استفاده از تابع رندم شافل[1] به صورت تصادفی در آورده، تا داده ها به صورت تصادفی در متغیر پخش شوند و مجموعه داده آزمون ترکیبی از کلاس های مختلف داده ما باشد.

۳-٤- تبدیل متن به بردار عددی با کتابخانه فست تکست

روش های مبتنی بر شبکه عصبی به دلیل دقت در زمینه پردازش زبان طبیعی بسیار رایج شده است. با این حال، اکثر این روش ها در تجزیه و تحلیل مجموعه داده های بزرگ زمان بر هستند و برای تجزیه و تحلیل متون نیاز به جاساز کلمه دارند. به همین دلیل، روشی به نام فست تکست پیشنهاد شد[١٤]. فست تکست کتابخانه ای است سریع، کارآمد و منبع باز که توسط تیم تحقیقاتی فیس بوک برای یادگیری کارآمد در نمایش کلمات و طبقه بندی جملات ایجاد شده است. در فست تکست، مجموعه ای از ترفندها برای بهبود سرعت پردازش و عملکرد مدل استفاده شده است که یکی از آنها اسکیپ گرام است. اسکیپ گرام اجازه می‌دهد برخی از کلمات حذف شوند و کلمات غیرمجاور با هم بررسی شوند[١٥]. فست تکست روشی برای تبدیل کلمات به بردار عددی است که فرمت آن همانند مدل ورد تو وک است. در کتابخانه فست تکست به جای یادگیری بردارها برای هر کلمه، هر کلمه را به صورت یک گرام از کاراکترها نشان می دهد. این کار برای درک معنای کلمات کوتاه تر کمک می کند و به جاساز ها اجازه می دهد پسوندها و پیشوندها را درک کنند. منظور از جاساز کلمات، این است که کلمات به صورت بردارهای کلمات متراکم ارائه می‌شوند، این به معنای این است که جاسازی کلمات اطلاعات بیشتری را در ابعاد کمتری گردآوری می‌کند. توجه به این نکته لازم است که جاسازهای کلمات، متن را به شکل قابل فهم برای انسان تبدیل نمی نماید، اما ساختار آماری زبان استفاده شده در نوشتار را نسبتا درک می‌کنند. هدف آن‌ها نگاشت معنایی به فضای هندسی است. این فضای هندسی فضای جاساز نامیده می‌شود. فست تکست با کلمات کمیاب به خوبی کار می کند. بنابراین اگر کلمه ای در طول آموزش دیده نشود، می توان آن را به ان گرام[2] هایی تقسیم کرد تا جاسازی شود. کتابخانه فست تکست شامل مدل های ورد تو وک و یک مدل از پیش آموزش دیده است که می تواند برای کارهایی مانند طبقه بندی جملات به خوبی استفاده شود. برای شناخت بهتر فست تکست با یک مثال چگونگی طبقه بندی یک جمله شرح داده می شود. برای مثال جمله " تصمیم گیری هر شخص بر اساس عقل است"، ابتدا توسط کتابخانه هضم کلماتی فاقد معنا مانند "هر" و "بر" حذف می شوند سپس توسط فست تکست به بخش های جداگانه شکست داده می شود.(شکل۳)

توکن های جمله

تصمیم	گیری	شخص	اساس	عقل	است

شکل(۳): بخش بندی جمله توسط فست تکست

کتابخانه فست تکست معنی هر یک از این بخش ها را جستجو خواهد کرد و مجموعه ای از اعداد (که بردار لغت نام دارد) را به هر لغت که در داده آموزشی حداقل یکبار ظاهر شده نسبت می دهد. این اعداد، معنی هر لغت را به عنوان یک نقطه در یک فضای ۳۰۰ بعدی، کد نگاری می شود.

شکل٤ نگاهی بر اعداد واقعی ازکتابخانه فست تکست که معنی هر کلمه را در جمله ما نمایش می دهند.

کلمه					
تصمیم	۰.۰۷۳۹۴۰۶۵	۰.۰۳۵۳۳۲۰۷	۰.۰۵۴۸۶۴۹۶	.....	-۰.۱۹۸۷۹۸۹۲
گیری	-۰.۰۲۶۲۷۸۲	۰.۰۳۲۸۹۷۷۳	۰.۰۱۳۰۸۲۱۸	.....	-۰.۰۸۵۲۲۴۳۵
شخص	۰.۲۴۹۶۱۹۴۲	۰.۲۶۰۲۱۷۳۷	۰.۱۵۲۷۷۶۶	.....	۰.۰۵۶۲۷۱۲۱
اساس	-۰.۳۶۸۵۶۴۹۶	-۰.۰۳۵۸۹۹۴۳	۰.۰۸۰۶۳۶۵۳	.....	-۰.۰۹۹۶۳۲۵۴
عقل	۰.۰۴۳۷۰۸۷۴	-۰.۱۵۴۴۵۴۰۸	-۰.۰۳۷۸۳۱۸۶	.....	۰.۰۸۴۲۱۷۹۶
است	۰.۲۶۴۴۸۶۷۶	۰.۰۶۵۱۷۱۳۸	۰.۱۸۶۴۱۰۸۴	.....	۰.۰۶۱۱۷۶۶

شکل(٤): اعدادی که فست تکست به هر کلمه در جمله می دهد

[1] Random shuffle

[2] N-gram

معنی هر لغت توسط ۳۰۰ ارزش، کدگذاری شده است. سپس، فست تکست از ستون های عمودی اعداد نمایش دهنده هر لغت میانگین می گیرد تا یک نمایش ۳۰۰ عددی از معنای کل جمله که بردار سند نام دارد، به وجود آورد (شکل۵).

جمله				
تصمیم گیری	۰.۰۳۹۴۸۵۴۰۲	۰.۰۳۳۸۷۷۵۰۷	۰.۰۷۴۹۸۹۸۷۵	.....	۰.۰۳۰۳۳۱۶۷۳
شخص اساس					
عقل است					

شکل(۵): نمایش ۳۰۰ عددی از معنای کل جمله

نهایتاً این ۳۰۰ ارزشی که جمله را نمایش می دهند، به عنوان یک ورودی از عنوان طبقه بندی می باشد.

در روش پیشنهادی، ماتریس ها طبق رابطه ۱الی ٤ مشخص می شوند. ماتریسی را با مقادیر صفر برای داده های آموزش ایجاد می شود و داده های آموزشی در این ماتریس قرار می گیرد. اندازه این ماتریس مطابق رابطه (۱) تعیین می‌شود.

train_size* max_no_tokens* vector_size (۱)

که در آن train_size برابر تعداد سطر های داده های آموزش، max_no_token برابر تعداد مجاز کلماتی که می‌تواند در یک جمله باشد و vector_size برابر با تعداد ستون های ماتریس است، از آنجایی که مدل فست تکست مورد استفاده برای ۳۰۰ ستون است، در این مقاله vector_size را برابر ۳۰۰ در نظر گرفته می شود.

در گام دوم همین روند برای ماتریس آزمون نیز انجام می‌شود. بدین صورت که ماتریسی با مقادیر صفر برای داده های آزمون ایجاد و داده های آزمون در این ماتریس قرار می گیرند. اندازه این ماتریس مطابق رابطه (۲) تعیین می شود.

test_size* max_no_tokens* vector_size (۲)

که در این رابطه test_size برابر با تعداد سطر های داده های آموزش، max_no_token برابر با تعداد مجاز کلماتی که می تواند در یک جمله باشد و vector_size برابر با تعداد ستون‌های ماتریس در نظر گرفته شده است.

ماتریسی با مقادیر صفر که برای عنوان داده‌های آموزش ایجاد می‌شود. مطابق رابطه (۳) و مقادیر آنها مقادیر پیش فرضی هستند که با توجه به تعداد عنوان(دسته) داده ها آموزش تعیین می شوند.

train_size*4 (۳)

ماتریسی با مقادیر صفر که برای عنوان داده های آزمون ایجاد می شود مطابق رابطه (٤) و مقادیر آنها مقادیر پیش فرضی هستند که با توجه به تعداد عنوان(دسته) داده ها آموزش تعیین می شوند.

test_size*4 (٤)

پس از ایجاد ماتریس های تمام صفر، نوبت به ایجاد بردار های عددی متون می رسد. در این مرحله به ازای هر کلمه ی جمله، یک بردار عددی توسط کتابخانه فست تکست ایجاد شده و در ماتریس آموزش یا ماتریس آزمون جای گذاری می شود.

۳-۵- ساخت مدل شبکه یادگیری عمیق

برای عملکرد بهتر سیستم نیاز به ساخت مدلی داریم که بهترین ویژگی‌های داده های ورودی را استخراج و به سیستم آموزش بدهد. در این مقاله از چندین لایه در ساخت مدل استفاده شده است.

۳-۵-۱- لایه شبکه کانولوشن[1] (شبکه عصبی پیچشی)

شبکه‌های عصبی کانولوشنال اخیراً عملکرد قابل ملاحظه ای در زمینه عملی برای طبقه بندی جملات به دست آورده اند. با این حال، در این روش باید یک مدل دقیق با پارامترهای همراه، از جمله اندازه ناحیه فیلتر، پارامترهای منظم سازی[2] و غیره تعیین شود[١٦]. معماری شبکه های پیچشی مشابه با الگوی اتصال نورون‌ها در مغز انسان است و از سازمان‌دهی قشر بصری در مغز الهام گرفته شده است. هر نورون به محرک‌ها تنها در منطقه محدودی از میدان بصری که تحت عنوان میدان تاثیر شناخته شده است پاسخ می‌دهد. یک مجموعه از این میدان‌ها برای پوشش دادن کل ناحیه بصری با یکدیگر هم‌پوشانی دارند.

در شکل (٦) معماری شبکه عصبی کانولوشن نشان داده شده است. شبکه کانولوشن معمولا از بلوک‌های مختلفی تشکیل شده است. لایه‌ها یا بلوک‌های مختلف در شبکه CNN عبارتنداز:

- لایه ورودی بلوک زرد

[1] Conventional Neural Network (CNN)

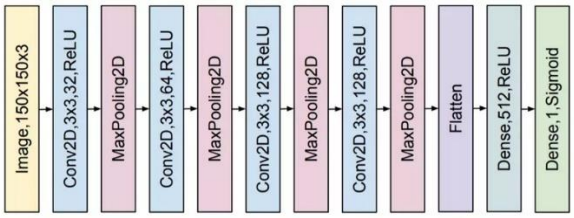
[2] Regularization

-لایه کانولوشن بلوک‌های آبی

-لایه غیرخطی لوک‌های آبی در شکل (معمولاً تابع غیرخطی را همراه با لایه کانولوشنی یک‌جا نشان می‌دهند)

-لایه ادغام[1] بلوک‌های قرمز

- لایه تمام متصل[2] بلوک‌های سبز



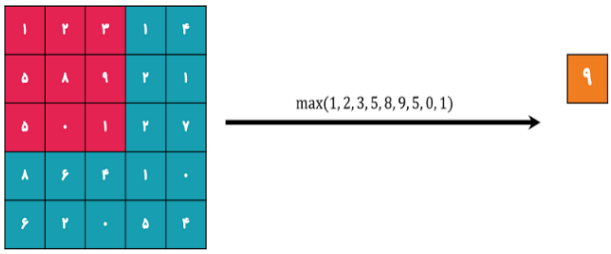
شکل(٦): معماری شبکه عصبی کانولوشن[۱۷]

در اولین مرحله روش پیشنهادی از یک لایه کانولوشن استفاده شده است. علت آن، این است که شبکه عصبی پیچشی به هر یک از اشیا و جنبه‌های موجود در متن میزان اهمیت (وزن‌های قابل یادگیری و بایاس) تخصیص می‌دهد و قادر به متمایزسازی آن‌ها از یکدیگر است. همچنین در این شبکه در مقایسه با دیگر الگوریتم‌های دسته‌بندی به پیش‌پردازش کمتری نیاز است. شبکه عصبی پیچشی، در صورتی که به اندازه کافی آموزش دیده باشد قادر به کسب فراگیری فیلترهای مشخصات است.

تعداد نورون هایی که در لایه کانولوشن در نظر میگیریم بر اساس آزمون و خطا و با توجه به بالاترین کارایی تعیین می کنیم. اندازه پنجره لغزان[3] را یک ماتریس ۵*۵ در نظر میگیریم. این اندازه می تواند باعث بازدهی بهتر مدل گردد. همچنین بین توابع فعال ساز، محاسبات در تابع ReLU ساده و صرفاً یک مقایسه است به همین خاطر محاسبات در بخش تابع غیرخطی با استفاده از ReLU نسبت به سایر توابع غیرخطی مانند سیگموید با سرعت بیشتری انجام می‌شود[۱۸]. از طرفی فرآیند آموزش با ReLU نسبت به سایر توابع غیرخطی سریع‌تر است. در نتیجه از تابع ReLU استفاده شده است. از آنجایی که استفاده از پنجره لغزان ابعاد ماتریس ورودی ما را کم می‌کند در نتیجه از لایه گذاری لبه ها[4] نیز استفاده شده تا به صورت جعلی اندازه ورودی را افزایش داده تا ماتریس خروجی هم‌اندازه ماتریس ورودی بشود.

۳-۵-۲ لایه بیشترین مقدار ادغام[5]

لایه ادغام یکی دیگر از لایه های مهم در شبکه عصبی کانولوشن است. هدف لایه ادغام کاهش اندازه مکانی نقشه ویژگی[6] بدست آمده با استفاده از لایه کانولوشنی است. لایه ادغام پارامتر قابل آموزش ندارد. صرفاً یک نمونه برداری ساده و موثر انجام می‌دهد. رایج‌ترین نمونه ادغام، بیشترین مقدار ادغام و میانگین ادغام[7] است . نتایج نشان می‌دهد که هیچ یک از این دو مدل برتری خاصی نسبت به مدل دیگر ندارد[۱۹]. مطابق شکل ۷ بیشترین مقدار ادغام شامل یک پنجره از پیش تعریف شده است (مثلا ۳×۳) که روی تصویر حرکت می‌کند (تصویر را اسکن می‌کند) و در هر پنجره مقدار بالاتر را انتخاب می‌کند و بقیه را دور می‌ریزد. مشابه با لایه کانولوشنی، اندازه فیلتر، لایه گذاری لبه ها و مقدار گام[8] در اینجا هم صدق می‌کند. در ادغام، معمولاً مقدار گام ۲ و اندازه فیلتر ۳ در نظر گرفته می‌شود. برای اینکه سایز ویژگی به 1/2 کاهش پیدا کند.



شکل(۷): مثالی از نحوه عملکرد لایه ادغام در شبکه عصبی کانولوشن

استفاده از بیشترین مقدار ادغام رایج تر است و عملکرد بهتری را از خود نشان داده است. پس از ادغام باید ماتریس تبدیل به یک بردار شده تا داده ها را به صورت ترتیبی به لایه متراکم[9] داده شود. برای این کار از لایه گلوبال ماکس پول[10] استفاده کرده و ماتریس خود را به یک بردار تبدیل می کنیم.

[1] Pooling layer
[2] Fully connected layer
[3] Sliding window
[4] Padding
[5] Max pooling
[6] Feature map

[7] Average pooling
[8] Stride
[9] Dense layer
[10] GlobalMaxPool1D

۳-۵-۳- لایه متراکم

با لایه‌ی متراکم تمامی گره‌ها در یک لایه به تمام گره‌ها در لایه بعدی متصل شده و معمولاً هر چه تعداد نورون‌ها در این لایه بیشتر شود صحت مدل نیز بیشتر می‌شود ولی از طرفی دیگر افزایش بیش از حد آن هم منجر به بیش پرازش[1] می‌شود. تابع فعال ساز استفاده شده در این لایه تابع بیشینه هموار[2] است، چرا که این تابع برای هریک از دسته ها یا عناوین یک عدد بین ۰ و ۱ تولید می نماید که با آن می توان میزان شباهت مقدار ورودی را با هریک از دسته ها بدست آورد.

۳-۵-٤- لایه بهینه ساز[3]

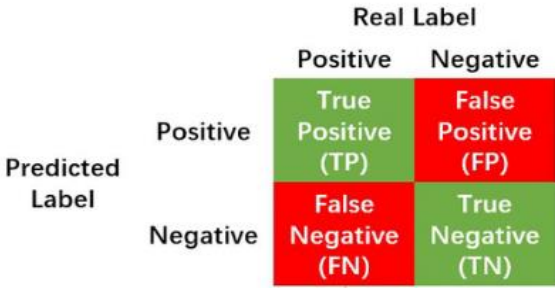
الگوریتم بهینه ساز فرآیند و نرخ یادگیری را کنترل می‌کند. بهینه‌سازها به الگوریتم‌ها یا روش‌هایی اطلاق می‌شود که برای تغییر ویژگی‌های شبکه عصبی از قبیل وزن و نرخ یادگیری به کار برده می‌شوند. این کار، نقش موثری در کاهش زیان دارد. نحوه تغییر وزن‌ها یا نرخ یادگیری شبکه عصبی به ابزارهای بهینه‌سازی مورد استفاده بستگی دارد. راهبردها یا الگوریتم‌های بهینه‌سازی مسئول کاهش زیان هستند و دقیق‌ترین نتایج ممکن را به ارمغان می‌آورند. بهینه ساز آدام[4] بسیار سریع است و به سرعت همگرایی پیدا می کند، همچنین نرخ یادگیری و واریانس بالا را اصلاح می کند[۲۰]. در روش پیشنهادی از این بهینه ساز استفاده شده است.

۳-٦- آموزش سیستم

در این مرحله، بر اساس مدلی که ساخته شده، داده های آموزشی به سیستم آموزش داده می شود. داده ها دو نوع هستند، نوع اول داده های اصلی و نوع دوم لیبل داده ها هستند که منتسب به داده های اصلی هستند. در این مرحله با توجه به پارامترهایی که وجود دارد، به سیستم آموزش داده می شود.

۳-٦-۱- ارزیابی مدل

در این مرحله از روش ارزیابی[5] مطابق رابطه (۵) برای اندازه‌گیری صحت داده مدل استفاده می شود.



شکل(۸): محاسبه میزان صحت داده ها

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (۵)$$

که در رابطه(۵)، TP بیانگر تعداد عنوان کلاس x است، که سیستم به درستی آن را x تشخیص داده است، FP. بیانگر تعداد عنوان های کلاس x است، که سیستم آن را x تشخیص نداده است، TN بیانگر تعداد عنوان های کلاسی به جز x که سیستم به درستی آن را x تشخیص نداده است، FN بیانگر تعداد عنوان های کلاسی به جز x است که سیستم آن را x تشخیص داده است.

۳-۷- میزان شباهت متن جدید به هر مکتب

در این مرحله، داده ای جدید از نویسنده دریافت شده و کلیه مراحل قبل از قبیل مرتب سازی متن، تبدیل متن به بردار عددی در این مرحله انجام و در نهایت میزان شباهت متن جدید با عنوان های آموزش داده شده به دست می آید.

٤- نتایج و آزمایشات

در این مقاله شش آزمایش مختلف با پارامترهای متفاوت انجام شد. جدول ۲ آزمایشات صورت گرفته در حالت‌های مختلف را به همراه میزان صحت حاصل شده از هر آزمایش نشان می‌دهد.

جدول ۲: میزان صحت داده در آزمایش های مختلف

شماره آزمایش	میزان صحت	پارامتر ها
۱	۷۸ درصد	استفاده از دو لایه کانولوشن
۲	۷۸ درصد	استفاده از ۲ لایه متراکم
۳	۷٦ درصد	استفاده از ٤ لایه متراکم
٤	۸٦ درصد	استفاده از تعداد نورون ۹۰۰ و ٤۵۰

[1] Overfitting
[2] Softmax function
[3] Optimizer
[4] Adam
[5] Evaluate

۵	۸۲ درصد	اندازه سایز بسته ۵۱۲
٦	۸۸ درصد	اندازه سایز بسته ۲۰٤۸

از آنجایی که لایه کانولوشن بیشتر در تصاویر پیچیده و چند بعدی استفاده می شود، لذا در پیاده سازی روش پیشنهادی که برای متون است، حالات استفاده از ایجاد یک لایه کانولوشنی و دو لایه کانولوشنی مورد مقایسه قرار گرفته و می توان نتیجه گرفت استفاده بیشتر از یک لایه کانولوشن در متون، باعث کاهش یادگیری می شود.

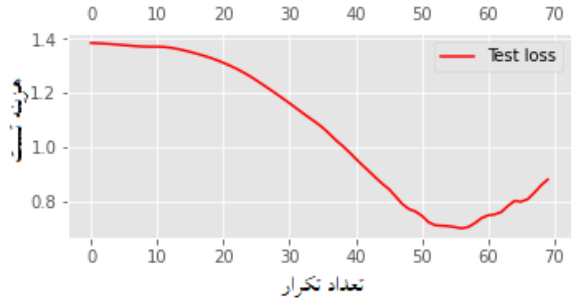
در روش پیشنهادی تعداد نورون لایه اول متراکم را برابر با ٦۰۰، تعداد نورون لایه دوم را ۳۰۰ و از آنجایی که تعداد دسته بندی ما ٤ است تعداد نورون های لایه سوم ٤ در نظر گرفته شده است .

مشخص کردن تعداد نورون ها بر اساس سعی و خطا به دست می آید. در آزمایش بررسی شده مشاهده می شود افزایش تعداد نورون بیشتر از ٦۰۰ باعث بیش پرازش شده، و در نتیجه تعداد نورون ایده آل در روش پیشنهادی برای لایه اول ٦۰۰ و برای لایه دوم ۳۰۰ است.

در روش پیشنهادی بالاترین یادگیری زمانی است که از ۳ لایه متراکم استفاده شده است. در زمان استفاده از ۲ لایه متراکم دقت صحت تا تکرار ۵۵، ۷۸ درصد و زمان استفاده از چهار لایه متراکم دقت صحت تا تکرار ٤۲، ۷۸ درصد می باشد. تکرار بیشتر از ٤۲ منجر به بیش پرازش می شود.

در این مرحله پارامتر هایی از قبیل تعداد دفعات یادگیری[1] تعیین اندازه اینکه داده ما به در قالب چند تایی به سیستم آموزش داده شود[2] و تقسیم بندی داده ها با هر بار تکرار، تعیین می شود.

پارامتر تعداد دفعات تکرار بر اساس سعی و خطا تعیین می شود و تا زمانی که یادگیری بیشتر و دچار بیش پرازش نشده قابل افزایش است. همینطور که در شکل ۹ قابل مشاهده است از تکرار ۵۵ به بعد ما دچار بیش پرازش شده، پس حداکثر تعداد تکرار ۵۵ در نظر گرفته می شود.

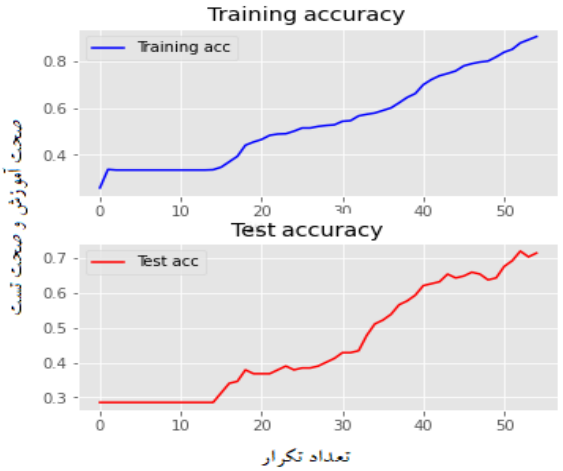


شکل(۹):حداکثر تعداد تکرار

پارامتر اندازه بسته در هر بار تکرار نقش مهمی در یادگیری دارد، اگر مقدار آن زیاد باشد یادگیری ما کم تر و اگر مقدار آن کم باشد بیش پرازش خواهیم داشت..زمانی که اندازه بسته ۵۱۲ در نظر گرفته شد، از تکرار ۳۰ به بعد دچار بیش پرازش شده، و میزان صحت داده ها تا تکرار ۳۰، ۸۳ درصد است و زمانی که اندازه بسته ۲۰٤۸ در نظر گرفته شد، تا تکرار ۵۵ میزان صحت داده ها، ۸۸ درصد است.

مقدار اندازه بسته که در ساخت مدل استفاده شده است برابر ۱۰۲٤ میباشد.(شکل ۱۰)

در مقاله با توجه به موضوع که برای طبقه بندی است برای تابع هزینه[3] از کراس آنتروپی[4] استفاده شده است.



شکل(۱۰): نمودار خروجی پس از ساخت مدل با روش پیشنهادی

[1] Epochs
[2] Batch size
[3] Cost function
[4] Crossentropy

در نهایت پس از آزمایش های متعدد صورت گرفته، پس از ساخت مدل بر اساس جدول ۳، میزان صحت داده های آموزش ۹۴ درصد می باشد.

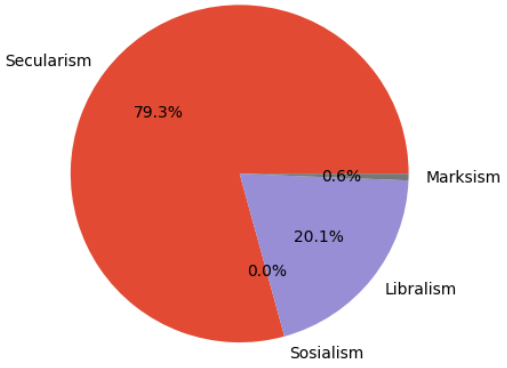
جدول(۳):لایه های استفاده شده در روش پیشنهادی

شماره	لایه های استفاده شده به ترتیب
۱	یک لایه کانولوشن
۲	یک لایه ادغام
۳	یک لایه گلوبال ماکس پول
٤	۲ لایه متراکم با تعداد نورون ٦۰۰ و ۳۰۰ و تابع فعالساز Relu
٥	۱ لایه متراکم با تعداد نورون ٤ و تابع فعال ساز بیشینه هموار
٦	یک لایه بهینه ساز آدام

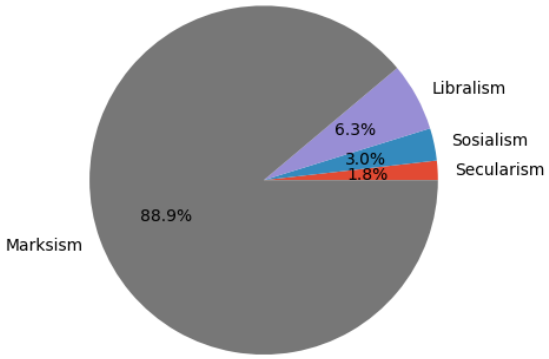
بعد از پایان آموزش مثال هایی (جدول ٤) از خروجی کار انجام شده که در شکل ۱۱، ۱۲، ۱۳ و ۱٤ قابل مشاهده است.

جدول(٤):مثال هایی از متن نویسنده و میزان شباهت با هر مکتب

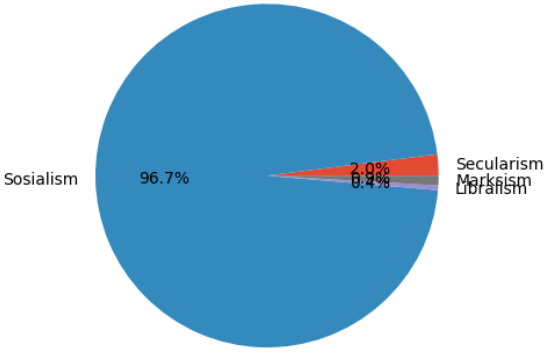
شماره مثال	متن استفاده شده
۱	جدایی دین از سیاست
۲	هر شخص باید به اندازه نیازش داشته باشد
۳	ابزار تولید باید دارای مالکیت دولت باشد
٤	تصمیم گیری هر شخص بر اساس عقل است



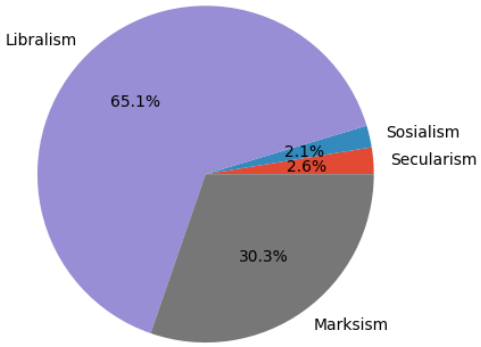
شکل(۱۱): متن "جدایی دین از سیاست"



شکل(۱۲): متن "هر شخص باید به اندازه نیازش داشته باشد"



شکل(۱۳): متن "ابزار تولید باید دارای مالکیت دولت باشد"



شکل(۱٤): متن "تصمیم گیری هر شخص بر اساس عقل است"

۵- نتیجه گیری و کارهای آینده

در این مقاله سعی شد تا با استفاده از شبکه های عصبی و یادگیری عمیق مکاتب فلسفی متعددی به سیستم آموزش داده شود و عنوان هایی را به داده ها اختصاص داده، همچنین با دادن داده جدید به سیستم، بتوان عنوان آن داده را مشخص کرد.

مرتب سازی متون قبل از عملیات یادگیری عمیق امری بسیار ضروری و مهم است، چرا که تعداد کلمات غیر ضروری بسیار بیشتر از کلمات مورد نیاز هستند و این خود هم از لحاظ افزایش یادگیری، هم از لحاظ سرعت یادگیری و هم از لحاظ مصرف حافظه دارای تاثیر می باشد. برای تبدیل متون به بردار های عددی، روش های متعددی وجود دارد، که در این مقاله از کتابخانه فست تکست استفاده شده است. کتابخانه فست تکست شامل مدل های ورد تو وک و یک مدل از پیش آموزش دیده است که می تواند برای کارهایی مانند طبقه بندی جملات به خوبی استفاده شود. یکی از موارد قابل توجه دیگر، انتخاب حداکثر تعداد کلمات مجاز در یک جمله است، زیرا در اندازه ماتریس و در نتیجه در سرعت و مصرف حافظه تاثیر گذار است. در روش پیشنهادی از یک لایه پیچشی استفاده شده است، زیرا استفاده بیشتر در موضوعات متنی باعث کاهش یادگیری می شود. برای استفاده از تابع فعال ساز، از تابع بیشینه هموار استفاده شده است. زیرا این تابع برای هریک از دسته ها یا عناوین یک عدد بین ۰ و ۱ می‌دهد که با آن می‌توان میزان شباهت مقدار ورودی را با هریک از دسته ها یا عناوین مشاهده نمود. تابع زیان در روش پیشنهادی را با توجه به اینکه موضوع یک پروژه دسته بندی چند کلاسه است از طبقه بندی کراس آنتروپی به عنوان پارامتر برای تابع زیان استفاده شده است. انتخاب مقدار اندازه بسته ها مهم می باشد، زیرا زیاد یا کم بودن آن کارایی یادگیری عمیق را پایین می آورد.

از جمله موضوعاتی که می توان در آینده مورد بحث و بررسی قرار بگیرد، امکان استفاده از تعداد عنوان های بیشتر مکاتب فلسفی و همچنین استفاده از مطالب با حجم های گسترده تر که در افزایش یادگیری بسیار موثر خواهد بود. می توان ساخت مدل با لایه های متفاوت برای پیاده سازی موضوعات مشابه و همچنین مقایسه با روش پیشنهادی انجام داد. همچنین اجرای روش های متفاوت برای تبدیل متن به بردارهای عددی با استفاده از کتابخانه های دیگر و مقایسه با روش پیشنهادی خواهد بود.

References

[۱] مطهری مرتضی، آشنائی با علوم اسلامی، ج۱، نشر صدرا، صفحه۱٤۷.

[۲] ابونصر فارابی، محمد بن محمد، الجمع بین رأیی الحکیمین، صفحه ۲۸.

[۳] ابن‌سینا، حسین بن عبدالله، الهیات شفا، صفحه ۱۵.

[4] Deepa Yogish, T. N. Manjunath , Ravindra S. Hegadi. “Review on Natural Language Processing Trends and Techniques Using NLTK”: 17 July 2019

[5] RyanSpring ; MatthewJohnson. “The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK and SpaCy tools”: June 2022

[6] Piotr Semberecki , Henryk Maciejewski , All Authors. “Deep learning methods for subject text classification of articles”: 3-6 Sept. 2017

[7] Mikolov, et al. “Efficient Estimation of Word Representations in Vector Space.” ArXiv.org, 7 Sept. 2013

[8] Tejas Menon. “Empirical Analysis of CBOW and Skip Gram NLP Models” Summer 2020

[9] Takamune Onishi; Hiromitsu Shiina. “Distributed Representation Computation Using CBOW Model and Skip–gram Model”: 1-15 Sept. 2020

[10] Lucky Agarwal , Kartik Thakral , Gaurav Bhatt , Ankush Mittal. “Authorship Clustering using TF-IDF weighted Word-Embeddings”: 12 December 2019

[11] Shervin Minaee , Nal Kalchbrenner , All Authors. “Deep Learning--based Text Classification: A Comprehensive Review” 17 April 2021

[12] Pranav Malik; Aditi Aggrawal; Dinesh K. “Toxic Speech Detection using Traditional Machine Learning Models and BERT and fastText Embedding with Deep Neural Networks”: April 2021

[13] Tengjun Yao; Zhengang Zhai. “Text Classification Model Based on fastText”: September 2020

[14] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov. “Bag of Tricks for Efficient Text Classification” ; [Submitted on 6

Jul 2016 (v1), last revised 9 Aug 2016 (this version, v3)]

[15] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space” ; [Submitted on 16 Jan 2013 (v1), last revised 7 Sep 2013 (this version, v3)]

[16] Ye Zhang, Byron Wallace. “A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification” ; [Submitted on 13 Oct 2015 (v1), last revised 6 Apr 2016 (this version, v4)]

[17] Marek Słoński, Krzysztof Schabowicz, Ewa Krawczyk “Detection of Flaws in Concrete Using Ultrasonic Tomography and Convolutional Neural Networks”; [27 March 2020]

[18] Dabal Pedamonti. “Comparison of non-linear activation functions for deep neural networks on MNIST classification task” ; [8 Apr 2018]

[19] FlorentinBieder, Robin Sandkühler, Philippe C. Cattin. “Comparison of Methods Generalizing Max- and Average-Pooling” ; 2 Mar 2021

[20] Diederik , P. Kingma,, Jimmy Ba. “Adam: A Method for Stochastic Optimization” ;[Submitted on 22 Dec 2014 (v1), last revised 30 Jan 2017 (this version, v9)]

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 1, No. 2, Autumn & Winter 2023, pp. 118-130

Identifying schools from authors' texts using data mining

Ahmad Alimohamadi[1], Mohamadmahdi Hoseini[2*]

[1]Computer Engineering Department, Shahrood Branch, Islamic Azad University, Shahrood, Iran

[2]Computer Engineering Department, Shahrood Branch, Islamic Azad University, Shahrood, Iran

Article Information

Original Research Paper

Received:
2023 June 25

Accepted:
2023 November 21

Keywords:
Philosophical schools, data mining of text, recognition of schools, conversion of text into a numerical vector, degree of similarity of author's thinking with schools

Corresponding Author*:
hosseini_mm@yahoo.com

Abstract

In the field of knowledge of philosophical schools, each person's thinking is different according to the type of attitude he can have regarding different schools. Recognizing the author's attitude and its similarity to each of the philosophical schools has always been one of the important issues in the field of humanities. In this article, a method based on deep learning is proposed to distinguish philosophical schools from text. In the proposed method, first the texts are normalized and redundant and meaningless words are removed. After the normalization stage, the text is broken into sentences and words, and then using the fasttext library, each word is converted into a numerical vector, after that, the features of the texts are extracted using the designed network, and finally, the system has learned and is ready to extract data, and by giving a new sentence, its similarity to each school is expressed. Based on the evaluation, the accuracy of the proposed method is 94%.

: 10.22034/ABMIR.2023.20250.1029

E-ISSN: 2821-2037 /© 2023. Published by Yazd University This is an open access article under the CC BY 4.0 License (https://creativecommons.org/licenses/by/4.0/).