ISLAMIC AZAD UNIVERSITY

SCIENCE AND RESEARCH BRACH

MASTER OF SCIENCE THESIS

COMPUTER ENGINEERING

– SOFTWARE

Subject:

Providing a Model for Detecting Personality of People in Social Networks.

Thesis Advisors:

Mehdi Hosseinzadeh

Peyman Sheikholharam mashhadi

By:

Omid Asghari

Autemn 2018

TABLE OF CONTENTS

Title **Page Number**

Chapter 3: Research Methdology

Chapter 4: Research Findings

Chapter 5 : Results and suggestions

Shapes index

Title **Page Number**

Abstract

The exponential growth of the use of social networks in cyberspace has led individuals to share a lot of information, including image, voice and text. Analyzing social Networks data provides details information about individual personality. The complexity and large volume of extracted data is that such that it requires to apply machine learning algorithms. In this paper the author has analyzed the behavior patterns using writing. we first need to know the standard personality models. one of the most reliable models is MBTI model. the goal of this thesis is to find a supervised Learning model that can determine personality factors by people writing in social networks. due to the fact that experience has shown for complex problems with many parameters, deep learning methods can be more effective, we used deep learning model and two personality factors are considered an introversion - extroversion and intuition - sensing. The obtained results show a good accuracy which based on we were able to distinguish an introversion - extroversion personality factor with precision of 62 % accuracy and intuition – sensing factor with precision of 58 %

Keywords: Social networks, MBTI, Deep Learning, supervised learning, Machine learnig.

Refrences:

[1] S. Fortunato, “Community detection in graphs,” *Phys. Rep.*, vol. 486, no. 3–5, pp. 75–174, 2010.

[2] L. Tang, X. Wang, and H. Liu, “Community detection via heterogeneous interaction analysis,” *Data Min. Knowl. Discov.*, vol. 25, no. 1, pp. 1–33, 2012.

[3] A. Rowe, S. Rowe, A. Silverman, and M. L. Borum, “P024 Crohn’S Disease Messaging on Twitter: Who’S Talking?,” *Gastroenterology*, vol. 154, no. 1, pp. S13–S14, 2018.

[4] B. Agarwal, “Personality Detection from Text : A Review,” *Int. J. Comput. Syst.*, vol. 1, no. 1, pp. 1–4, 2014.

[5] Y. Amichai-Hamburger, “Internet and personality,” *Comput. Human Behav.*, vol. 18, no. 1, pp. 1–10, 2002.

[6] R. Ackoff, “Ackoff’s Best,” pp. 170–172, 1999.

[7] R. Samizade, E. Mahmoudi, and S. Abad, “The Application of Machine Learning Algorithms for Text Mining based on Sentiment Analysis Approach,” *J. J. Inf. Technol. Manag.*, vol. 10, no. 102, pp. 309–330, 2018.

[8] E. Younesi, L. Toldo, B. Müller, C. M. Friedrich, N. Novac, A. Scheer, M. Hofmann-Apitius, and J. Fluck, “Mining biomarker information in biomedical literature,” *BMC Med. Inform. Decis. Mak.*, vol. 12, no. 1, 2012.

[9] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, “Predicting Personality from Twitter.pdf,” 2011.

[10] C. C. Doi, “Our Twitter Profiles,Our Selves: Predicting Personality with Twitter,” pp. 180–185, 2011.

[11] R. Wald, T. Khoshgoftaar, and C. Sumner, “Machine prediction of personality from Facebook profiles,” *Proc. 2012 IEEE 13th Int. Conf. Inf. Reuse Integr. IRI 2012*, pp. 109–115, 2012.

[12] S. Poria, A. Gelbukh, and B. Agarwal, “Advances in Soft Computing and Its Applications,” vol. 8266, no. November, 2013.

[13] B. Verhoeven, W. Daelemans, and T. De Smedt, “Ensemble Methods for Personality Recognition,” *Proc. Work. Comput. Personal. Recognit. Shar. Task*, pp. 1–4, 2013.

[14] A. Ma and G. Liu, “Neural Networks in Predicting Myers Brigg Personality Type From Writing Style,” pp. 1–9, 2017.

[15] A. Essazadegan and D. Ph, “ Relationship between the metaphors and Eysenck's introversion-extraversion dimensions,” pp. 54–63,.

معاونت پژوهش و فن آوری

به نام خدا

منشور اخلاق پژوهش

یاری از خداوند سبحان و اعتقاد به این که عالم محضر خداست و همواره ناظر بر اعمال انسان و به منظور پاس داشت مقام بلند دانش و پژوهش و نظر به اهمیت جایگاه دانشگاه در اعتلای فرهنگ و تمدن بشری، ما دانشجویان و اعضاء هیات علمی واحدهای دانشگاه آزاد اسلامی متعهد می گردیم اصول زیر را در انجام فعالیت های پژوهشی مد نظر قرار داده و از آن تخطی نکنیم:

۱- اصل حقیقت جویی: تلاش در راستای پی جویی حقیقت و وفاداری به آن و دوری از هرگونه پنهان سازی حقیقت.

۲- اصل رعایت حقوق: التزام به رعایت کامل حقوق پژوهشگران و پژوهیدگان (انسان، حیوان و نبات) و سایر صاحبان حق.

۳- اصل مالکیت مادی و معنوی: تعهد به رعایت کامل حقوق مادی و معنوی دانشگاه و کلیه همکاران پژوهش.

۴- اصل منافع ملی: تعهد به رعایت مصالح ملی و در نظر داشتن پیشبرد و توسعه کشور در کلیه مراحل پژوهش.

۵- اصل رعایت انصاف و امانت: تعهد به اجتناب از هرگونه جانب داری غیر علمی و حفاظت از اموال، تجهیزات و منابع در اختیار.

۶- اصل رازداری: تعهد به صیانت از اسرار و اطلاعات محرمانه افراد، سازمان ها و کشور و کلیه افراد و نهادهای مرتبط با تحقیق.

۷- اصل احترام: تعهد به رعایت حریم ها و حرمت ها در انجام تحقیقات و رعایت جانب نقد و خودداری از هرگونه حرمت شکنی.

۸- اصل ترویج: تعهد به رواج دانش و اشاعه نتایج تحقیقات و انتقال آن به همکاران علمی و دانشجویان به غیر از مواردی که منع قانونی دارد.

۹- اصل برائت: التزام به برائت جویی از هرگونه رفتار غیر حرفه ای و اعلام موضع نسبت به کسانی که حوزه علم و پژوهش را به شائبه های غیر علمی می آلایند.

دانشگاه آزاد اسلامی

واحد علوم تحقیقات تهران

تعهدنامه اصالت رساله یا پایان‌نامه

اینجانب امید اصغری دانش‌آموخته مقطع کارشناسی‌ارشد ناپیوسته در رشته مهندسی کامپیوتر گرایش نرم افزار که در تاریخ 1397/10/29 از پایان نامه خود تحت عنوان " ارائه‌ی مدلی برای تشخیص شخصیت افراد بر اساس متن منتشر شده در شبکه‌های اجتماعی." با کسب نمره 17/50 و درجه خوب دفاع نموده‌ام بدینوسیله متعهد می‌شوم :

1) این پایان‌نامه حاصل تحقیق و پژوهش انجام شده توسط اینجانب بوده و در مواردی که از دستاوردهای علمی و پژوهشی دیگران (اعم از پایان‌نامه، کتاب، مقاله و ...) استفاده نموده‌ام، مطابق ضوابط و رویه موجود، نام منبع مورد استفاده و سایر مشخصات آن را در فهرست مربوطه ذکر و درج کرده‌ام.
2) این پایان‌نامه قبلا برای دریافت هیچ مدرک تحصیلی (هم سطح، پایین‌تر یا بالاتر) در سایر دانشگاه‌ها و مؤسسات آموزش عالی ارائه نشده است.
3) چنانچه بعد از فراغت از تحصیل، قصد استفاده و هرگونه بهره‌برداری اعم از چاپ کتاب، ثبت اختراع و ... از این پایان‌نامه داشته باشم، از حوزه معاونت پژوهشی واحد مجوزهای مربوطه را اخذ نمایم.
4) چنانچه در هر مقطع زمانی خلاف موارد فوق ثابت شود، عواقب ناشی از آن را می‌پذیرم و واحد دانشگاهی مجاز است با اینجانب مطابق ضوابط و مقررات رفتار نموده و در صورت ابطال مدرک تحصیلی‌ام هیچ‌گونه ادعایی نخواهم داشت.

نام و نام‌خانوادگی :

امید اصغری

تاریخ و امضاء :

1397/10/29

دانشگاه آزاد اسلامی

واحد علوم تحقیقات تهران

دانشکده مکانیک، برق و کامپیوتر، گروه مهندسی کامپیوتر

پایان‌نامه برای دریافت درجه کارشناسی‌ارشد در رشته مهندسی کامپیوتر **(M.Sc)**

گرایش: نرم افزار

عنوان:

ارائه‌ی مدلی برای تشخیص شخصیت افراد بر اساس متن منتشر شده در شبکه‌های اجتماعی

استادان راهنما:

دکتر مهدی حسین زاده

دکتر پیمان شیخ‌الحرم مشهدی

نگارش:

امید اصغری

پاییز 1397

سپاس خداوندی را که همه چیزم از اوست

خدایا

تو را که به من زندگی دوباره بخشیدی

تو را را که به من هر لحظه امید می بخشی

تو را که هدفم را در جهت درست پیش می رانی

ای خدای من برای تمامی اینها سپاس می گویمت

با سپاس و تشکر فراوان از اساتید راهنمای ارجمندم **جناب آقای دکتر مهدی حسین زاده و جناب آقای دکتر پیمان شیخ الحرم مشهدی** که بی دریغ و صبورانه مرا در به سرانجام رساندن این تحقیق یاری نمودند.

با تقدیر و درود فراوان خدمت خانواده ی دلسوز و فداکارم

پدر و مادرم

برادر و عموی دلسوزم که در تمامی مراحل زندگی ام بهترین مشوقان من بوده اند و پیوسته جرعه نوش جام تعلیم و تربیت، فضیلت و انسانیت آنها بوده ام و همواره چراغ وجودشان روشنگر راه من در سختی ها و مشکلات بوده است.

با سپاس بی دریغ از دوستان گران مایه ام مخصوصا جناب آقای دکتر علیرضا ابوحسین

که مرا صمیمانه و مشفقانه یاری داده اند.

و با تشکر خالصانه خدمت همه ی کسانی که مرا در به انجام رساندن این امر مهم یاری نموده اند.

احمقانه است که کسی موفقیت هایش را تمام و کمال مخصوص خویش بداند، همواره دست ها، قلب ها و افکار بسیاری در موفقیت های ما سهیم هستند.

(والت دیسنی)

فهرست مطالب

عنوان **شماره صفحه**

فصل سوم: روش اجرایی تحقیق

فصل چهارم: یافته‌های پژوهش

فصل پنجم: نتایج و پیشنهادات

فهرست اشکال

عنوان *شماره صفحه*

چکیده:

گسترش شگرف اینترنت و استفاده روزافزون از شبکه‌های اجتماعی موجب شده‌است که افراد روزانه اطلاعات زیادی را به اشتراک بگذارند، این اطلاعات شامل تصویر، صدا، متن و غیره می‌باشد. با تحلیل شبکه‌های اجتماعی می‌توان اطلاعات گوناگونی استخراج کرد اما گاهی تحلیل این شبکه‌ها به دلیل وجود داده‌های عظیم و پیچیدگی مسئله به دست انسان ممکن نیست بنابراین الگوریتم‌های یادگیری ماشین را برای تحلیل آنها به کار می‌بریم. این الگوریتم‌ها برای کشف و یادگیری الگوهای موجود در شبکه‌های اجتماعی به کار می‌روند مانند الگوهای رفتاری، ارتباطی و غیره. الگوهای رفتاری‌ای که اشخاص از آن‌ها پیروی می‌کنند می‌تواند ناشی از شخصیت درونشان باشد، این الگوهای رفتاری شامل حرف زدن و نوشتار هم می‌شود به این معنی که اشخاص با شخصیت‌های مختلف نوشتار متفاوتی دارند. ما برای این کار ابتدا باید مدل‌های شخصیتی استاندارد را بشناسیم، یکی از معتبرترین مدل‌ها، مدل MBTI است. هدف این تحقیق ارائه‌ی مدلی است که بتواند به صورت یادگیری با نظارت بر اساس نوشتار افراد در شبکه‌های اجتماعی فاکتورهای شخصیتی آنان را بر اساس مدل شخصیتی MBTI حدس بزند. تجربه نشان داده است برای مسائل پیچیده با تعداد پارامترهای زیاد روش‌های یادگیری عمیق می‌توانند موثرتر واقع شوند، در این تحقیق برای تشخیص الگوها از روش یادگیری عمیق استفاده شد و دو فاکتور شخصیتی درونگرا – برونگرا و شمی – حسی بررسی شد. نتایج بدست آمده نشان دهنده‌ی دقتی مناسب است که بر اساس آن فاکتور شخصیتی درونگرا – برونگرا با دقتی معادل 62٪ و فاکتور شخصیتی شمی – حسی را با دقتی معادل 58٪ تشخیص داده شد.

کلمات کلیدی: شبکه‌های اجتماعی، MBTI، یادگیری عمیق، یادگیری با نظارت، یادگیری ماشین.

فصل اول:

کلیات تحقیق

1-1- مقدمه

امروزه به دلیل حضور اینترنت، شبکه‌های اجتماعی به عنوان بخشی جدایی ناپذیر از زندگی بشر محسوب می شوند. ما درعصر شبکه‌ها زندگی می‌کنیم . عصری که در آن شکل‌گیری شبکه‌های گوناگون اجتماعی آنلاین، شیوه‌های ارتباطی واطلاع‌رسانی نوینی به عرصه گسترده ارتباطات اجتماعی معرفی کرده‌است. براساس آمار مؤسسهٔ استیستا[1] یک میلیارد و 790 میلیون نفر از مردم جهان در شبکه های اجتماعی اینترنتی عضو هستند. نرخ نفوذ شبکه های اجتماعی 31 در صد ا ست و 73 در صد از مردم ایالت متحده عضو یکی از شبکه‌های اجتماعی هستند. بررسی شبکه‌های اجتماعی که زمان چندانی هم از پیدایش آنها نمی‌گذرد، موضوع مورد علاقه بسیاری از دانشجویان و پژوهشگران حوزه علوم ارتباطات اجتماعی است. کاربرد این شبکه‌های نوظهور در طیف وسیعی از روابط شخصی گرفته تا روابط جهانی، این پدیده را تبدیل به یک سوژه تمام عیار در پژوهش‌های اجتماعی و رسانه‌ای کرده‌است. با کمی دقت در شبکه‌های اجتماعی می‌توان اطلاعات گوناگونی ا ستخراج کرد. گاهی تحلیل این شبکه‌ها به دلیل وجود داده‌ی عظیم و پیچیدگی مسئله به دست انسان ممکن نیست بنابراین الگوریتم‌هایی برای تحلیل آنها به کار می‌بریم. این الگوریتم‌ها برای کشف و یادگیری الگوهای موجود در شبکه‌های اجتماعی به‌کار می‌روند. الگوهای رفتاری، ارتباطی و غیره.

[1]statista

1-2- بیان مسئله

یک شبکه اجتماعی ساختاری اجتماعی است، متشکل از گره‌هایی است که به صورت منفرد یا گروهی(سازمانی) مرتبط شده‌اند و روابطی همچون مبادلات مالی، دوستی، بازرگانی، پیوندهای وب و سرگرمی دارند[1]. چنین شبکه‌هایی مردم را با انواع علایق مختلف به یکدیگر مرتبط می‌کند و به مردم اجازه می‌دهند که در مورد مسائل مختلف صحبت کنند[2].

امروزه رشد و تغییر نمایی استفاده از شبکه‌های اجتماعی مجازی موجب شده‌است که افراد روزانه اطلاعات زیادی به اشتراک بگذارند. این اطلاعات شامل تصویر، صدا، متن و غیره می‌باشد. شبکه‌های اجتماعی یکی از بهترین منابع برای تحلیل و آنالیز پدیده‌های مختلف محسوب می‌شوند[3]. یک چالش بحث برانگیز محققین امروز این است که آیا این داده‌ها می‌تواند حاوی اطلاعاتی در مورد شخصی که آنها را به اشتراک گذاشته‌است یا شند؟ الگوهای رفتاری‌ای که اشخاص از آنها پیروی می‌کنند نا شی از شخصیت درونشان است[4]. این الگوهای رفتاری شامل حرف زدن و نوشتار هم می‌شود. به این معنی که اشخاص با شخصیت‌های مختلف نوشتار مختلفی دارند. ما برای این کار ابتدا باید مدل‌های شخصیتی استاندارد را بشناسیم. یکی از معتبرترین مدل‌ها، مدل MBTI است. پدیدآورندگان این مدل کاترین کوک بریگز[1] و دخترش، ایزابل بریگز مایرز[2] بودند که بر اساس تحقیقاتشان روی مطالعات یونگ این آزمون را طراحی کردند. آن‌ها این آزمون را ابتدا در طی جنگ جهانی دوم بدین منظور ارائه کردند تا بتوانند مناسب‌ترین شغل را برای زنانی که در صنعت نظامی کار می‌کردند پیدا کنند. بر اساس این مدل افراد در 16 کلاس شخصیتی تقسیم می‌شوند. این مدل برای شخصیت 4 فاکتور را در نظر می‌گیرد که عبارتند از:

درونگرا- برونگرا

شمی – حسی

منطقی-احساسی

قضاوتی- ادراکی

[1]Katharine Cook Briggs

2Isabel Briggs Myers

هر فرد در آن واحد می‌تواند از هر فاکتور شـــخصـــیتی در یک گروه قرار بگیرد. برای مثال یک فرد می‌تواند درونگرا، شمی، احساسی، قضاوتی باشد.این فاکتورها در نوشتار تاثیر می‌گذارند. هدف این پایان‌نامه یافتن مدل و الگوریتمی است که بتواند بر اساس نوشتار افراد در شبکه‌های اجتماعی فاکتورهای شخصیتی آنان را حدس بزند. بنابراین ابتدا باید کلمات را برای کامپیوتر طوری تعریف کنیم که معنی آنها را نیز دربرداشـــته باشد. تا با استفاده از آن بتواند جملات را تفسیر و کلاس‌بندی کند.

1-3- اهمیت و ضرورت تحقیق

وجود حجم عظیم داده در فضای مجازی محققان را به استفاده و تحلیل این داده‌ها سوق داد. از این این داده‌ها اطلاعات زیادی می‌توان استخراج کرد. امروزه مردم در شبکه‌های مجازی نظرات، افکار، رویدادها و اتفاقات روزمره‌ی خود را بیان می‌کنند. برای تصمیمات و برنامه‌های در سطح یک جامعه که نیاز به بازخورد مردم دارند از این اطلاعات می‌توان استفاده کرد. بنابراین با استفاده از این داده‌ها می‌توان نظرات یک جامعه را در مورد یک رویداد، یک محصول یک شخص و غیره تحلیل کرد.

هنگامی که یک محصول وارد بازار می‌شود، سازندگان آن برای به‌روزرسانی نسخه‌های بعدی و کسب درآمد بیشتر یا رسیدن به هدف‌های تعریف شده برای آن محصول نیاز به بازخورد کاربران دارند تا محصول را متناسب با نیاز آنان به‌روزرسانی کنند. یک راهکار ساده برای این مسئله که اجرا می شود قرار دادن مکانی برای دریافت نظرات مردم اســت، اما داده‌ای که از طریق شـبکه‌های مجازی بدســت می‌آید بسـیار بزرگ تر است و بنابراین از این طریق می‌توان چشم‌انداز بهتری از نظرات کاربران داشت.

شرکت‌ها و یا افرادی که می‌خواهند از محصول، نظر، ایده و یا یک رویداد تبلیغ کنند با دانستن شخصیت افرادی که در فضای مجازی حضور دارند، می‌تواند تبلیغات خود را متناسب با شخصیت هر فرد برای او بفرستند.

در همین راستا شناخت شخصیت و هویت در فضای مجازی به یک چالش بزرگ تبدیل شده است. بر اساس [5] شخصیت عامل اصلی نحوه‌ی رفتار افراد در اینترنت است. از آنجا که شبکه به دلیل ماهیت خود متشـکل از تعاملات انسـانی اسـت، نتیجه می‌شـود که ما نمی‌توانیم نحوه ی عملکرد اینترنت را بدون درک شخصیت افرادی که در آن تعامل می‌کنند بفهمیم.

درک شخصیت افراد توسط کامپیوتر نیاز به مدلی دارد که در آن کلمات و جملات مفهوم داشته باشند و ســـپس کامپیوتر با روش‌های یادگیری ماشـــین بتواند این مدل را حین فرایند یاد بگیرد. ما از الگوریتم‌های یادگیری نظارت شده برای ارائه‌ی بهترین مدل استفاده خواهیم کرد.

ورودی این الگوریتم توئیت‌ها و یا هر پستی که افراد به صورت متنی در فضای مجازی منتشر کرده‌اند بعلاوه تیپ شخصیتی آن‌ها می‌باشد و مدل بایستی این ورودی‌ها را یاد بگیرد.

1-4- اهداف

1-4-1- هدف کلی

هدف اصلی این پژوهش آنالیز و پیش‌بینی رفتار افراد در فضای مجازی و هویدا کردن شخصیت واقعی افراد و رابطه‌ی بین رفتار کاربران در اینترنت و شخصیت آنها توسط ماشین است. همچنین در دست داشتن الگوریتمی که بتواند شخصیت افراد را بر اساس گفتار و نوشتارشان طبق مدل شخصیتی MBTI که جزئیات بیشتری را از شخصیت افراد در اختیار می‌گذارد، با دقت مناسبی تشخیص دهد.

1-4-2- هدف فرعی

تعریف و ارائه‌ی کلمات به کامپیوتر بصورت معنادار و درک رابطه‌ی بین کلمات و جملات از طریق الگوریتم‌ها و اعداد توسط کامپیوتر از اهداف این پایان نامه می‌باشد همچنین کاوش در داده‌های متنی کاربران جهت آشنا شدن با شخصیت آنها نیز از دیگر اهداف این تحقیق است.

1-5- سوالات تحقیق

در این تحقیق سوالاتی مطرح می‌گردد که در انتهای کار بایستی به آنها پاسخ دهیم. برخی از این سوالات به شرح زیر می‌باشند:

- آیا ماشین می‌تواند به شخصیت افراد پی ببرد؟
- آیا رابطه‌ای بین شخصیت افراد و رفتارشان در اینترنت وجود دارد؟
- آیا برای تعامل با کاربران جهت استخراج الگوهای شخصیتی آنها تکنولوژی‌های جدیدی وجود دارد؟
- چه مدلی برای پیش بینی شخصیت افراد بر اساس برخی از فاکتورهای آزمون MBTI وجود دارد؟

1-6- فرض و فرضیات تحقیق

در طراحی مدل پیشنهادی این پایان‌نامه فرض‌هایی در نظر گرفته شده است که عبارتند از :

- الگوریتمی وجود دارد که شخصیت افراد را از روی گفتارشان تشخیص دهد.
- پست‌هایی که افراد در صفحات اجتماعی به اشتراک می‌گذارند می‌تواند اطلاعات و موضوعاتی را درباره آنها آشکار سازد.
- استخراج اطلاعات در مورد شخصیت افراد در شبکه‌های مجازی توسط ماشین امکان پذیر است.
- با استفاده از اطلاعات پروفایل یک شخص در صفحات مجازی می‌توان شخصیت او را پیش بینی کرد.
- رابطه‌ای بین اطلاعات نوشتار افراد و نوع شخصیت آنها وجود دارد.

1-7- ساختار پایان‌نامه

در این تحقیق اطلاعات اولیه که به صورت داده ی متنی شامل تیپ شخصیتی افراد و 50 عدد از پستهایشان به زبان انگلیسی می‌باشد از سایت داده کاوی www.kaggle.com گرفته شده‌است.

این مجموعه داده در این سایت برای یک مسابقه‌ی انتخاب بهترین مدل تشخیص شخصیتی قرار داده شده است. از این افراد اطلاعات دیگری در مجموعه داده وجود ندارد. برای مدل MBTI این مجموعه داده بزرگترین مجموعه داده ی موجود است. ما برای هر فرد 50 پست از او را داریم هدف کلاس بندی این متون است بگونه‌ای که برچسب کلاس با برچسبی که برای شخصیت فرد در نظر گرفته شده است برابر باشد.

٪80 داده را برای آموزش سیستم استفاده می‌کنیم و آن را برای ساختن مدل به کار می‌بریم. یعنی از 8600 نمونه حدودا 7000 نمونه برای آموزش و باقی برای ارزیابی و محاسبه‌ی خطای مدل به کار می‌رود. هدف به حداقل رساندن خطای مدل بر روی مجموعه داده‌ی ارزیابی است.

متدهای هوشمند زیادی برای حل این مسئله استفاده شده‌است که شامل متدهای هوش مصنوعی کلاسیک و پیشرفته می‌شود. برای مثال می‌توان از SVM ، MLP و Deep Neural Networks نام برد. تجربه نشان داده‌است برای مسائل کلاس‌بندی با تعداد متغیر و ابعاد بالا به دلیل وجود پارامترهای یادگیرنده‌ی زیاد، متدهای یادگیری عمیق نتایج به نسبت بهتری را نشان داده‌اند.

ابتدا با مطالعه منابع کتابخانه‌ای به بررسی روند کلی کار، کارهای انجام شده، بررسی چارچوب‌ها و الگوریتم های موجود در زمینه آنالیز متن می‌پردازیم.

فرآیند کلی ساختن مدل شامل سه مرحله است:

۱.آماده سازی داده‌ها (پیش‌پردازش داده)

۲.تفسیر داده‌ها برای کامپیوتر

۳.ساخت و ارزیابی مدل

در فاز اول ما داده‌ها را در اصطلاح تمیز می‌کنیم. نویزها را برطرف میکنیم اطلاعات غیر مفید شامل آدرس وبسایت هایی که پست افراد در آنها بوده‌است را حذف می‌کنیم. توزیع داده‌ها را در کلاس‌ها را برای آموزش یکسان می‌کنیم و در واقع داده‌ها را برای ورود به سیستم آماده می‌کنیم.

در مرحله‌ی دوم فایل داده‌ای که پیش پردازش شده است را باید بخوانیم و کلمات آن را با بردارهای عددی برای مدل نمایش دهیم. در این مرحله از الگوریتم‌های نمایش کلمات مانند glove، Word2Vec استفاده می‌کنیم و هر کلمه را به صورت بردار به قسمی نمایش می‌دهیم که کلماتی که هم معنی هستند و یا معانی نزدیک‌تر و مرتبط‌تری به هم دارند بردارهایشان به هم نزدیک‌تر باشند. همچنین کلماتی که بیشتر با هم ممکن است در یک متن ظاهر شوند نیز بردارهای نزدیک به همی داشته باشند، در این مرحله باید تعیین کنیم که

طول این بردارها چقدر باید باشد که این یک فراپارامتر است و بسته به شرایط مسئله ما در اینجا طول بردارها را 50 انتخاب کرده‌ایم. در مرحله‌ی سوم مدل را طراحی می‌کنیم و بردارهای کلماتی را که آماده کرده بودیم برای کلاس‌بندی به مدل می‌دهیم. تا الگوی مورد نظر را برای این مسئله با کمترین میزان خطای ممکن بیابیم.

فصل دوم :

ادبیات تحقیق

2-1- مقدمه

در بخش مبانی نظری و ادبیات تحقیق ، به‌منظور روشنگری مفاهیم نظری و هویت مفهوم، با برشمردن ویژگی‌ها و ابعاد آن، اقدام به ارائه این بخش می‌نماییم. در این بخش قصد داریم به توضیح و تشریح اصلاحات و چهارچوب‌های تحقیق بپردازیم.

2-2- مبانی نظری و ادبیات تحقیق

2-2-1- پردازش زبان‌های طبیعی

پردازش زبان‌های طبیعی یکی از زیرشاخه‌های بااهمیت در حوزه‌ی گسترده علوم رایانه و هوش مصنوعی است که به تعامل بین کامپیوتر و زبان‌های (طبیعی) انسانی می‌پردازد؛ بنابراین پردازش زبان‌های طبیعی بر ارتباط انسان و رایانه، متمرکز است. پس چالش اصلی و عمده در این زمینه درک زبان طبیعی و ماشینی کردن فرایند درک و برداشت مفاهیم بیان‌شده با یک زبان طبیعیِ انسانی است . به تعریف دقیق‌تر، پردازش زبان‌های طبیعی عبارت است از استفاده از رایانه برای پردازش زبان گفتاری و زبان نوشتاری. بدین معنی که رایانه‌ها را قادر سازیم که گفتار یا نوشتار تولید شده در قالب و ساختار یک زبان طبیعی را تحلیل و درک نموده یا آن را تولید نمایند. در این صورت، با استفاده از آن می‌توان به ترجمه زبان‌ها پرداخت، از صفحات وب و بانک‌های اطلاعاتیِ نوشتاری جهت پاسخ دادن به پرسش‌ها استفاده کرد، یا با دستگاه‌ها، مثلاً برای مشورت گرفتن به گفت‌وگو پرداخت.

این‌ها تنها مثال‌هایی از کاربردهای متنوع پردازش زبان‌های طبیعی هستند. گفتنی است هنوز سامانه‌ی کارآمدی برای پردازش زبان‌های طبیعی به‌وجود نیامده‌است.هدف اصلی در پردازش زبان طبیعی، ایجاد تئوری‌هایی محاسباتی از زبان، با استفاده از الگوریتم‌ها و ساختارهای داده‌ای موجود در علوم رایانه است. بدیهی است که در راستای تحقق این هدف، نیاز به دانشی وسیع از زبان است و علاوه بر محققان علوم رایانه، نیاز به دانش زبان شناسان نیز در این حوزه می‌باشد. با پردازش اطلاعات زبانی می‌توان آمار مورد نیاز برای کار با زبان طبیعی را استخراج کرد. کاربردهای پردازش زبان طبیعی به دو دسته کلی قابل تقسیم است: کاربردهای نوشتاری و کاربردهای گفتاری. از کاربردهای نوشتاری آن می‌توان به استخراج اطلاعاتی خاص از یک متن، ترجمه یک متن به زبانی دیگر یا یافتن مستنداتی خاص در یک پایگاه داده نوشتاری اشاره کرد. نمونه‌هایی از کاربردهای گفتاری پردازش زبان عبارتند از: سیستم‌های پرسش و پاسخ انسان با رایانه، سرویس‌های اتوماتیک ارتباط با مشتری از طریق تلفن، سیستم‌های آموزش به فراگیران یا سیستم‌های کنترلی توسط صدا. در سالهای اخیر این حوزه تحقیقاتی توجه دانشمندان را به خود جلب کرده‌است و تحقیقات قابل ملاحظه‌ای در این زمینه صورت گرفته‌است.

2-2-1-1- داده، اطلاعات، دانش

طبق نظر راسل ایکاف که یک تئوریسین سیستمی و پرفسور تغییر سازمانی می باشد محتوی ذهن بشر به پنج دسته می‌تواند طبقه‌بندی شود:

۱- داده‌ها[1]: سمبل‌ها، داده‌ها از منابع حیاتی به شمار می روند.

۲- اطلاعات[2]: داده‌های پردازش شده که می توانند مفید و مورد استفاده واقع شوند یا به عبارتی می توان گفت با بهره‌برداری صحیح از داده‌ها می توان داده‌ها را به اطلاعات بامعنی تبدیل کرد و پاسخ به سوال های درباره چه کسی، کجا، چه وقت و چه را فراهم می‌کنند .

۳- دانش[3]: کاربرد و استفاده از داده‌ها و اطلاعات می باشند و پاسخ به سوالاتی درباره چگونگی را فراهم می‌سازند.

۴- فهم[4]: کاربرد چرا.

۵- خرد[5]: ارزیابی فهم و درک.

راسل ایکاف در [6] به این مسئله اشاره می کند که ۴ دسته اول (داده – اطلاعات- دانش – فهم) به گذشته مرتبط هستند و با این مسئله که چه بوده است و چگونه شناخته شده است سرو کار دارند و فقط مقوله پنجم که خرد می باشد با آینده در ارتباط است زیرا با تلفیق بینش و دانش طراحی شده است .

درک و فهم : یک فرایند متغیر و احتمالی است یا به عبارت دیگر یک فرایند شناختی و تحلیلی است می توانیم دانش را بگیریم و با دانش های اندوخته‌شده از قبل ترکیب کنیم .تفاوت بین درک و دانش تفاوت میان یادگیری و به خاطر سپردن است .افرادی که درک می کنند می توانند اعمال وفعالیتهای مفیدی انجام دهند زیرا آنها می توانند دانش جدید را یا در برخی موارد حداقل اطلاعات جدید را از آنچه قبلا می دانسته اند ترکیب کنند . بدین طریق است که درک می تواند برروی اطلاعات تازه ساخته شده ، دانش و درک بنایی جدید بسازد . در زبان کامپیوتر سیستم هوش مصنوعی دارای درک به مفهومی که قادر است دانش جدید را با اطلاعات و دانش از پیش ذخیره شده را ترکیب کند.

خرد: یک فرایند غیر قطعی و غیر احتمالی است و گسترده و وسیع می باشد و مخصوصا نوع خاص برنامه ریزی بشری (اخلاقی ،رمزهای وابسته به علم اخلاق و..) خرد جوهره کاوش فلسفی است برخلاف ۴ سطح قبلی سوالاتی درباره آنچه که پاسخی ندارد (به آسانی قابل تحقق نیست) می پرسد و در برخی اوقات به مواردی می پردازد که دوره زمانی برای پاسخ شناخته شده بشری نمی تواند وجود داشته باشد .بنابراین خرد فرایندی است

[1] Data
[2] Information
[3] Knowledge
[4] Understanding
[5] Wisdom

که ما می توانیم بین خوب و بد ، غلط و درست را تشخیص دهیم یا قضاوت کنیم و ایکاف معتقد است که کامپیوتر نمی تواند این کار را انجام بدهد.

داده: یک حقیقت یا بیان حادثه را بدون ارتباط به چیزهای دیگر ارائه می کند .

اطلاعات : درک ارتباط از یک گونه ، احتمالا علت و معلول را در بر می گیرد .برای مثال دما تا ۱۵ درجه افت می کند و سپس باران می آید .

دانش : یک الگو ارائه می دهد که معمولا یک سطح بالاتر از پیش بینی است به طوری که از آنچه شرح داده شده است یا در آینده اتفاق خواهد افتاد ، مرتبط است وآن را تهیه می کند.

خرد : بیش از فهم اصول بنیادی موجود در دانش را در بر می گیرد .خرد شامل چیزی بیشتر از درک اصول بنیادی دانش است که این اصول خود پایه و اساس دانش را تشکیل می دهد .خرد اساسا سیستماتیک است.

2-2-1-2- مراحل فرایند کشف دانش از پایگاه داده‌ها

فرایند کشف دانش از پایگاه داده ها شامل پنج مرحله است که عبارتند از :

1. انبارش داده‌ها
2. انتخاب داده‌ها
3. تبدیل داده‌ها
4. کاوش در داده‌ها
5. تفسیر نتیجه

همانگونه که مشاهده می شود داده کاوی یکی از مراحل این فرایند است که به عنوان بخش چهارم آن نقش مهمی در کشف دانش از داده ها ایفا می کند .

داده‌ها:

وجود اطلاعات صحیح و منسجم یکی از ملزوماتی است که در داده کاوی به آن نیازمندیم . اشتباه و عدم وجود اطلاعات صحیح باعث نتیجه گیری غلط و در نتیجه اخذ تصمیمات ناصحیح در سازمان‌ها می گردد و منتج به نتایج خطرناکی خواهد گردید که نمونه‌های آن کم نیستند .

اکثر سازمان‌ها دچار یک خلا اطلاعاتی هستند . در اینگونه سازمان‌ها معمولا سیستم‌های اطلاعاتی در طول زمان و با معماری و مدیریت‌های گوناگون ساخته شده‌اند ، به طوری که در سازمان اطلاعاتی یکپارچه و مشخصی مشاهده نمی‌گردد . علاوه بر این برای فرایند داده کاوی به اطلاعات خلاصه و مهم در زمینه تصمیم گیری‌های حیاتی نیازمندیم .

هدف از فرایند انبارش داده‌ها فراهم کردن یک محیط یکپارچه جهت پردازش اطلاعات است . در این فرایند، اطلاعات تحلیلی و موجز در دوره‌های مناسب زمانی سازماندهی و ذخیره می‌شود تا بتوان از آنها در فرایندهای

تصمیم گیری که از ملزومات آن داده‌کاوی است ،استفاده شود . به طور کلی تعریف زیر برای انبار داده ها ارائه می گردد :

انبار داده‌ها ، مجموعه ای است موضوعی، مجتمع ، متغیر در زمان و پایدار از داده‌ها که به منظور پشتیبانی از فرایند مدیریت تصمیم گیری مورد استفاده قرار می گیرد .

انبارش داده‌ها خود موضوع مفصلی است که مقاله‌ها و رساله‌های گوناگونی در مورد آن نگاشته شده اند . در این فصل به منظور آشنایی با این فرایند به آن اشاره‌ای شد .

انتخاب داده‌ها:

انبار داده‌ها شامل انواع مختلف و گوناگونی از داده‌ها است که همه آنها در داده‌کاوی مورد نیاز نیستند . برای فرایند داده‌کاوی باید داده‌ها ی مورد نیاز انتخاب شوند . حتی در مواردی نیاز به کاوش در تمام محتویات پایگاه نیست بلکه ممکن است به منظور کاهش هزینه عملیات ، نمونه‌هایی از عناصر انتخاب و کاوش شوند.

تبدیل داده‌ها

هنگامی که داده‌های مورد نیاز انتخاب شدند و داده‌های مورد کاوش مشخص گردیدند ، معمولا به تبدیلات خاصی روی داده‌ها نیاز است . نوع تبدیل به عملیات و تکنیک داده‌کاوی مورد استفاده بستگی دارد : تبدیلاتی ساده همچون تبدیل نوع داده‌ای به نوع دیگر تا تبدیلات پیچیده تر همچون تعریف صفات جدید با انجام عملیات‌های ریاضی و منطقی روی صفات موجود .

کاوش در داده‌ها:

داده‌های تبدیل شده با استفاده از تکنیک‌ها و عملیات‌های داده‌کاوی مورد کاوش قرار می‌گیرند تا الگوهای مورد نظر کشف شوند .

تفسیر نتیجه:

اطلاعات استخراج شده با توجه به هدف کاربر تجزیه و تحلیل و بهترین نتایج معین می گردند . هدف از این مرحله تنها ارائه نتیجه (بصورت منطقی و یا نموداری) نیست ، بلکه پالایش اطلاعات ارایه‌شده به کاربر نیز از اهداف مهم این مرحله است .

2-2-1-3- عملیات داده‌کاوی

در داده‌کاوی، چهار عمل اصلی انجام می‌شود که عبارتند از:

1. مدلسازی پیشگویی کننده

2. تقطیع پایگاه داده‌ها
3. تحلیل پیوند
4. تشخیص انحراف

از عملیات اصلی مذکور ، یک یا بیش از یکی از آنها در پیاده سازی کاربردهای گوناگون داده‌کاوی استفاده می‌شوند . به عنوان مثال برای کاربرد های خرده فروشی معمولا از عملیات تقطیع و تحلیل پیوند استفاده می شود در حالی که برای تشخیص کلاهبرداری ، می‌توان از هر یک از چهار عملیات مذکور استفاده نمود . علاوه برا ین می توان از دنباله ای از عملیات برای یک منظور خاص استفاده کرد . مثلا برای شناسایی مشتریان، ابتدا پایگاه تقطیع می شود و سپس مدلسازی پیشگویی کننده در قطعات ایجاد شده اعمال می گردد .

تکنیک‌ها ، روش‌ها و الگوریتم‌های داده‌کاوی ، راه‌های پیاده سازی عملیات داده‌کاوی هستند . اگر چه هر عملیات نقاط ضعف و قوت خود را دارد ، ابزارهای گوناگون داده‌کاوی عملیات‌ها را بر اساس معیارهای خاصی، انتخاب می‌کنند . این معیارها عبارتند از :

- تناسب با نوع داده های ورودی
- شفافیت خروجی داده کاوی
- مقاومت در مقابل اشتباه در مقادیر داده ها
- میزان صحت خروجی
- توانایی کار کردن با حجم بالای داده ها

مدلسازی پیشگویی‌کننده:

مدلسازی پیشگویی کننده ، شبیه تجربه یادگیری انسان در به کار بردن مشاهدات برای ایجاد یک مدل از خصوصیات مهم پدیده‌ها است . در این روش از تعمیم دنیای واقعی و قابلیت تطبیق داده های جدید با یک قالب کلی، استفاده می شود .

در این مدل، می توان با تحلیل یک پایگاه داده‌های موجود ، خصوصیات مجموعه‌های داده را تعیین کرد . این مدل با استفاده از روش یادگیری نظارت شده، شامل دو فاز آموزش و آزمایش ایجاد شده‌است . در فاز آموزش با استفاده از نمونه‌های عظیمی از داده‌های سابقه‌ای ، مدلی ساخته می شود که به آن مجموعه آموزشی می گویند . در فاز آزمایش این مدل روی داده‌هایی که در مجموعه آموزشی قرار ندارند ، اعمال می شود تا صحت و خصوصیات آن تایید گردد .

تقطیع پایگاه داده‌ها:

هدف از تقطیع پایگاه داده‌ها ، تقسیم آن به تعداد نامعینی از قطعات یا خوشه‌هایی از رکوردهای مشابه است، یعنی رکوردهایی که خصوصیاتی مشابه دارند و می توان آنها را همگن فرض کرد. پیوستگی داخلی این قطعات بسیار زیاد است در حالی که همبستگی خارجی میان آنها کم می‌باشد .

در این مدل بر خلاف مدل قبل، از یادگیری نظارت نشده برای تعیین زیرشاخه‌های ممکن از جمعیت داده‌ای استفاده می شود. دقت تقطیع پایگاه داده‌ها از روش‌های دیگر کمتر است، بنابراین در مقابل خصوصیات نامربوط و افزونگی، حساسیت کمتری از خود نشان می‌دهد.

تحلیل پیوند:

در این روش پیوندهایی مرسوم به بستگی میان رکوردها و یا مجموعه‌ای از رکوردها بازشناسی می‌شوند . سه رده ویژه از تحلیل پیوند وجود دارند که عبارتند از :

1. کشف بستگی
2. کشف الگوهای متوالی
3. کشف دنباله‌های زمانی مشابه

تشخیص انحراف:

داده‌کاوی فرآیندی است که طی آن با استفاده از انواع مختلف ابزار تحلیل داده به دنبال کشف الگوها و ارتباطات میان داده‌های موجود که ممکن است منجر به استخراج اطلاعات جدیدی از پایگاه داده گردند می باشد.

اولین وساده ترین گام تحلیل داده در داده‌کاوی توضیح و شرح مشخص داده (از جمله معنی داده وانحراف استاندارد کلمه)می باشد که این کار می تواند به وسیله نمودارها و گراف ها وهمچنین کلماتی که با این کلمه ارتباط معنایی نزدیکی دارند انجام گردد در نتیجه جمع‌آوری، جستجو و انتخاب داده درست در این بخش بسیار مهم و حیاتی می باشد.

اما این کار به تنهایی کار خاصی انجام نمی‌دهد شما باید یک مدل پیش بینی کننده بر اساس الگوهایی که از نتایج دانش به دست آورده شده بسازید سپس آزمایش کنید که آیا آن مدل با نمونه اصلی سازگار است. یک مدل خوب نباید با جهان واقع تفاوت چندانی داشته باشد.

آخرین گام نیز تشخیص صحت وسقم عملکرد مدل بصورت تجربی می باشد. برای مثال از یک بانک مربوط به مشتریان وپاسخ هایی که به یک پیشنهاد خاص داده‌اند یک مدل می سازیم که بر اساس آن مشخص می

شود که کدام حدس وانتظار بیشترین نزدیکی را با یک پیشنهاد مانند پیشنهاد قبلی دارد و اینکه آیا می توانیم بر این حدس اعتماد کنید یا نه؟

2-2-1-3-1- الگوریتم‌های داده‌کاوی

حال برخی از الگوریتمها و مدلهایی را که برای کاوش داده استفاده می شود بررسی می‌کنیم. اغلب محصولات از انواع گوناگونی از الگوریتم‌ها که در علم کامپیوتر یا مقالات آماری ارائه شده به همراه پیاده سازی خاص آنها که جهت رسیدن به هدف فروشنده می‌باشد استفاده می‌نمایند. برای مثال بسیاری از فروشندگان نسخه-هایی از درختهای تصمیم CART یا CHAID را به همراه امکاناتی برای کار بر روی کامپیوترهای موازی می فروشند. برخی از فروشندگان الگوریتم‌های مختص خود دارند که گرچه ممکن است وابستگی‌ها یا امکانات اضافی نداشته باشد اما می‌تواند خوب کار کند.

شاید مهمترین نکته این باشد که هیچ مدل یا الگوریتمی نمی تواند و نباید به تنهایی استفاده شود. برای هر مساله داده شده طبیعت داده استفاده شده بر روی انتخاب مدل‌ها و الگوریتم‌هایی که بر می گزینیم تاثیر خواهد گذاشت. نمی‌توان هیچ مدل یا الگوریتمی را در این زمینه بهترین نامید. نتیجتا به یک سری ابزار و تکنولوژی جهت یافتن بهترین مدل ممکنه نیاز خواهیم داشت.

2-2-1-4- آنالیز متن و متن‌کاوی

متن‌کاوی، به داده‌کاوی‌ای که بر روی متن انجام شود اشاره دارد .همچنین به عنوان آنالیز متن نیز شناخته می‌شود که منظور از آن فرایند استخراج اطلاعات با کیفیت از متن است. اطلاعات پر کیفیت، بطور معمول از فهم الگوها و گرایشها از طریق معانی و بوسیله یادگیری الگوهای آماری حاصل می‌شود. متن کاوی معمولاً درگیر در فرایند ساختاردهی به ورودی‌های متنی (معمولاً تجزیه، همراه با افزودن برخی ویژگی‌ها تفاسیر زبانی و حذف موارد اضافی و درج موارد بعدی در پایگاه داده انجام می‌گیرد)، استخراج الگوهای درون داده‌های ساختار یافته، و در نهایت ارزیابی و تفسیر خروجی‌ها است. «پر کیفیت» در متن کاوی معمولاً به ترکیبی از مرتبط بودن، نو ظهور بودن و جالب بودن اشاره دارد. وظایف متن کاوی معمول شامل دسته‌بندی متون، خوشه بندی متون، استخراج معنی و مفهوم، تولید رده‌بندی دانه‌ای، تجزیه و تحلیل احساسات، خلاصه کردن اسناد و مدلسازی ارتباط موجودیت‌ها است. (بطور مثال یادگیری ارتباط بین موجودیتها).

اصطلاح آنالیز متن یک مجموعه از تکنیک‌های زبان‌شناسی، آمار و یادگیری ماشینی را توضیح می‌دهد که محتوای اطلاعات منابع متنی را برای هوشمند سازی کسب و کار، آنالیز اکتشافی داده، تحقیقها یا سرمایه‌گذاری ساختار داده و مدل می‌کند. این اصطلاح تقریباً مترادف متن کاوی است. اصطلاح آنالیز متن بیشتر در

کسب‌وکار مورد استفاده قرار می‌گیرد در حالی که متن کاوی حوزه کاربرهای قدیمیتر بویژه تحقیقها علوم وابسته به زندگی و هوشمند سازی دولت‌ها استفاده می‌شود .

اصطلاح آنالیز متن همچنان شرح می‌دهد که کاربرد آنالیز متن برای پاسخ به مشکل‌های کسب و کار، چه وابسته یا مستقل از پرس و جو و آنالیزهای میدانی و داده‌های عددی باشد. واضح است که ۸۰ درصد از اطلاعات وابسته به کسب و کار در شکلی بدون ساختار و متنی است. این تکنیک‌ها و فرایندها دانشی – حقایق، قواعد کسب و کار و ارتباطات – را کشف و ارائه می‌نمایند که در غیر این صورت در ساختاری متنی، غیرقابل نفوذ برای فرایندهای خودکار باقی مانده بودند .

2-2-1-4-1- کاربردهای آنالیز متن

یکی از رویکردهای نوین داده‌کاوی که به استخراج دانش از حجم وسیعی از داده‌های متنی می‌پردازد، آنالیز احساس نام دارد. آنالیز احساس نوعی زمینه‌ی تحقیقاتی است که به تحلیل نظرها،احساس‌ها، ارزیابی‌ها، رفتارها، گرایش‌ها، و عاطفه‌هاب بیان شده با یک زبان نوشتاری می‌پردازد. آنالیز احساس به استخراج احساسات و عقاید کاربران از متن‌های منتشر شده در صفحات اینترنتی کمک شایانی می‌کند[7].

عقاید دیگران در موقع تصمیم‌گیری یا انتخاب یک گزینه از میان چندین گزینه می‌تواند بسیار حیاتی باشد. حوزه مورد مطالعه تحلیل نظرات و احساسات مردم نسبت به مسائلی همچون محصولات، خدمات، سازمان‌ها، افراد، مسائل و رویدادها می‌باشد. بنابراین امروزه با گسترش روزافزون رسانه‌های اجتماعی (مثل بلاگ‌ها، توییتر و فیس‌بوک) افراد و سازمان‌ها از محتوای این رسانه‌ها برای تصمیم‌گیری استفاده می‌کنند و دیگر لازم نیست افراد به نظرات دوستان و آشنایان درباره یک محصول اکتفا کنند، زیرا عقاید و بحث‌های فراوانی درباره آن محصول در محیط وب وجود دارد. البته به خاطر فراوانی و تنوع سایت‌ها، پیداکردن و پایش سایت‌های مفید و چکیده گرفتن از اطلاعات موجود در آنها هنوز کاری دشوار و پیچیده می‌باشد، زیرا هر سایت به طور معمول دارای حجم عظیمی از نظرات می‌باشد که برای انسان عادی استخراج و خلاصه‌سازی نظرات موجود سخت می‌باشد. بنابراین به یک سیستم تشخیص نظرات و احساسات نیاز می‌شود. در سال‌های اخیر سیستم‌های تشخیص نظرات و احساسات، تقریبا در تمام محدوده‌های ممکن، از پیش‌بینی فروش محصولات و خدمات مشتریان گرفته تا پیش‌بینی وقایع اجتماعی و نتیجه انتخابات سیاسی توسعه یافته است.

2-2-2- کاربردهای تحلیل شبکه‌های اجتماعی:

یکی از کاربردهای تحلیل شبکه های اجتماعی در تصمیم گیری و خط مشی گذاری است. تحلیل شبکه های اجتماعی و شیوه ارتباطات مردم، می‌تواند کمک کند که سیاست گذاران، در زمینه تصمیم گیری برای اعلام یا عدم اعلام عمومی انواع بیماری ها، انتخاب بهتری انجام دهند. شرکتهای ارائه دهنده خدمات شبکه های

اجتماعی دیجیتال هم از جمله استفاده کنندگان جدی تحلیل شبکه های اجتماعی هستند . پیشنهاد ارتباطات جدید در فیس‌بوک و یا پیشنهاد تصاویر توسط اینستاگرام که ما به صورت روزانه با آنها سر و کار داریم، خروجی الگوریتم های تحلیل شبکه های اجتماعی است.

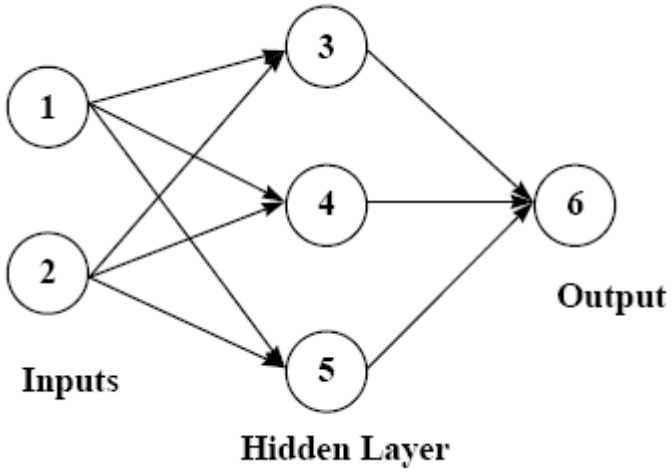
همچنین، گوگل را می‌توان یکی از بزرگترین متخصصان تحلیل شبکه های اجتماعی دانست . برای تحلیل شبکه های اجتماعی، پارامترها و شاخصهای متعددی طراحی شده و مورد استفاده قرار می‌گیرد.

میزان تعامل هر گره (Node) با گره های دیگر، تفاوت یا تشابه توزیع جغرافیایی گره ها و توزیع دیجیتال آنها، عمق و شدت نفوذ اثر هر رفتار گره بر روی گره های دیگر، قرار گرفتن هر گره در مرکز یا میزان فاصله گرفتن آن از مرکز شبکه، تعامل یکسویه یا دوسویه‌ی گره با سایر گره ها، تنوع اطلاعات و ارتباطات بین گره ها و انتروپی موجود در شبکه اجتماعی، از جمله صدها پارامتر و معیاری هستند که در تحلیل شبکه‌های اجتماعی مورد توجه قرار می‌گیرند.

2-2-3- شبکه‌های عصبی:

شبکه های عصبی به طور خاصی مورد استفاده اند چرا که آنها ابزاری موثر برای مدلسازی مسائل بزرگ و پیچیده که ممکن است در آنها صدها متغیر پیش بینی کننده که فعل و انفعالات زیادی دارند وجود داشته باشد.(شبکه های عصبی زیستی بطور غیر قابل مقایسه ای پیچیده تر هستند.)شبکه های عصبی می توانند در مسائل طبقه بندی یا حدسهای بازگشتی(که در آنها متغیر خروجی پیوسته است) استفاده شوند.

یک شبکه عصبی با یک لایه داخلی شروع می شود که در آن هر گره به یک متغیر پیشگو منسوب می گردد. این گره های ورودی به یک تعداد از گره ها در لایه پنهان متصل می شوند.گره ها در لایه پنهان می توانند به گره هایی در یک لایه پنهان دیگر یا به یک لایه خروجی متصل شود. لایه خروجی خود شامل یک یا بیشتر متغیرهای جواب می باشد.



شکل2-1- یک شبکه عصبی مصنوعی با سه لایه

2-2-4- یادگیری عمیق:

یادگیری عمیق به بیانی دیگر یادگیری ژرف ، (یادگیری ساختار ژرف یا یادگیری سلسله مراتبی) یک زیر شاخه از یادگیری ماشینی و بر مبنای مجموعه‌ای از الگوریتم‌ها است که در تلاش هستند مفاهیم انتزاعی سطح بالا در دادگان را مدل نمایند که این فرایند را با استفاده از یک گراف عمیق که دارای چندین لایه پردازشی متشکل از چندین لایه تبدیلات خطی و غیر خطی هستند، مدل می‌کنند. به بیان دیگر پایه آن بر یادگیری نمایش دانش و ویژگی‌ها در لایه‌های مدل است . یک نمونه آموزشی می‌تواند به صورت‌های گوناگون بسان یک بردار ریاضی پر شده از مقدار به ازای هر پیکسل و در دید کلی تر به شکل یک مجموعه از زیرشکل‌های کوچک‌تر مدل سازی شود. برخی از این روش‌های مدل سازی سبب ساده شدن فرایند یادگیری ماشین شده‌است. در یادگیری ژرف امید به جایگزینی استخراج این ویژگی‌های تصویر به دست بشر با روش‌های کامل خودکار بی‌نظارت و نیمه نظارتی وجود دار. انگیزه نخستین در بوجود آمدن این ساختار یادگیری از راه بررسی ساختار عصبی در مغز انسان الهام گرفته شده‌است که در آن یاخته‌های عصبی با فرستادن پیام به یکدیگر درک را امکان‌پذیر می‌کنند. بسته به فرض‌های گوناگون در مورد نحوهٔ اتصال این یاخته‌های عصبی، مدل‌ها و ساختارهای مختلفی در این حوزه پیشنهاد و بررسی شده‌اند، هرچند که این مدل‌ها به صورت طبیعی در مغز انسان وجود ندارد و مغز انسان پیچیدگی‌های بیشتری را دارا است. این مدل‌ها نظیر شبکه عصبی عمیق، شبکه عصبی پیچیده ، شبکه باور عمیق پیشرفت‌های خوبی را در حوزه‌های پردازش زبان‌های طبیعی، پردازش تصویر ایجاد کرده‌اند .

در حقیقت عبارت یادگیری عمیق، بررسی روش‌های تازه برای شبکه عصبی مصنوعی است(لی گومز،2014)[1].

2-2-5- تاریخچه:

با رشد فناوری اطلاعات و روش‌های تولید و جمع آوری داده ها، پایگاه داده های مربوط به داده های تبادلات تجار ی، کشاورزی، اینترنت، جزئیات مکالمات تلفنی، داده های پزشکی و غیره سریعتر از هر روز جمع آوری و انبارش می شوند. لذا از اواخر دهه 80 میلادی بشر به فکر دست یابی به اطلاعات نهفته دراین پایگاه داده های حجیم افتاد زیرا سیستمهای سنتی قادر به این کار نبودند. به دلیل رقابت در عرصه های سیاسی، نظامی، اقتصادی و علمی و اهمیت دست یابی به اطلاعات در کمترین زمان بدون دخالت انسان علم و تجزیه و تحلیل داده ها یا داده کاوی پا به عرصه گذاشت.

داده کاوی فرآیند ی است که در آغاز ده 90 مطرح شد و با نگرشی نو، به مسئلۀ استخراج اطلاعات از پایگاه داده ها می‌پردازد. از سال 1995 داده کاوی به صورت جدی وارد مباحث آمار شد و در سال 1996 ، اولین شمارۀ مجلۀ کشف دانش و معرفت از پایگاه داده ها منتشر شد . محققانی نظیر براچمن و آناند (1996) کلیه مراحل واقع گرا یانه و رو به جلو کشف دانش از پایگاه داده ها را تشخیص دادند.

در حال حاضر، داده کاوی مهمترین فناوری جهت بهره برداری موثر از داده های حجیم است و اهمیت آن رو به فزونی است . به طوریکه تخمین زده شده است که مقدار داده ها در جهان هر 20ماه به حدود دو برابر می رسد. در یک تحقیق که بر روی گرو ه های تجاری بسیار بزرگ در جمع آوری داده ها صورت گرفت مشخص گردید که 19 درصد از این گرو ه ها دارای پایگاه داده هایی با سطح بیشتر از 50 گیگا بایت می باشند و 59 درصد از آ نها انتظار دارند که در آینده ای نزدیک در چنین سطحی قرار گیرند .

درصنایعی مانند کارت‌های اعتباری و ارتباطات و فرشگاه های زنجیره‌ای و خریدهای الکترونیکی و اسکنرهای بارکد خوان هر روزه داده های زیادی تولید و ذخیره می شوند . افزایش سرعت کامپیوترها باعث به وجود آمدن الگوریتم‌هایی شده است که قدرت تجزیه و تحلیل بسیار بالایی دارند بدون اینکه محدودیتی در زمینه ظرفیت و سرعت کامپیوترها داشته باشند .

در سال 1989 و 1991 کارگا ههای کشف دانش و معرفت از پایگاه داده ها توسط پیاتتسکی[2] و همکارانش برگزار شد. در فواصل سا لهای 1991 تا 1994کارگا ههای کشف دانش و معرفت از پایگاه داده ها توسط فییاد[3]و پیاتتسکی و دیگران برگزار شد. به طور رسمی اصطلاح داده کاوی برای اولین بار توسط فییاد در اولین

[1] Lee Gomes
[2] Piatetsky
[3] Fayyad

کنفرانس بین المللی"کشف معرفت و داده کاوی[1] "در سال 1995 مطرح شد. امروزه کنفرانسهای مختلفی در این زمینه در سراسر دنیا برگزار میشود.

افزایش داده های بسیار باعث پیدایش فرصتهای تازه برای کار در علوم مهندسی و کسب و کار شده است. زمینه داده کاوی و کشف دانش از پایگاه داده ها به عنوان یک رشته علمی جدید در مهندسی و علوم کامپیوتر ظهور کرده است. مهندسی صنایع با حوزه های گوناگون و در بر داشتن فرصتهای بینظیر اکنون برای کاربرد داده کاوی و کشف دانش از پایگاه داده ها و برای توسعه مفاهیم و روشهای تازه در این زمینه آماده است. فرآیندهای صنعتی زیادی اکنون برای مطمئن شدن از کیفیت سفارشات محصول و کاهش هزینه های محصول به طور خودکار و کامپیوتری شده اند.

2-2-6- بازنمایی کلمات[2]

سامانه‌های پردازش صدا و تصویری که دقت بالایی دارند با استفاده از مجموعه دادگانی غنی و با ابعاد بالا کار می‌کنند که در آن‌ها تصاویر به‌صورت بردارهای از شدت نورهای پیکسل‌های خام و اصوات به‌صورت ضرایبی از شدت توان، کدبندی‌شده‌اند. برای کارهایی مانند تشخیص گفتار، ما می‌دانیم که تمامی دانش لازم برای انجام موفق این کار، در همان کدگشایی دادگان خام است (چون انسان این کارها را به‌خوبی از دادگان خام انجام می‌دهد).

سامانه‌های پردازش زبان طبیعی سنتی با لغات به‌صورت نمادهای گسسته‌ی اتمی (غیرقابل تجزیه) رفتار می‌کنند، مثلاً گربه می‌تواند به شکل Id537 و سگ به‌صورت Id143 بازنمایی شود. این کد سازی‌ها به صورت دلخواه است و هیچ‌گونه اطلاعات مفیدی که ممکن است در بین این لغات وجود داشته باشد را فراهم نمی‌سازد. به همین دلیل پژوهشگران درحال‌توسعه مدل‌هایی هستند که این روابط حاکم بین واژگان را در بازنمایی آن‌ها تعبیه سازد.

Word2vec یک مدل پیشگو به منظور یادگیری تعبیه سازی لغت از متن خام است که از لحاظ پیچیدگی محاسباتی بسیار ساده است .به بیان ساده تر در این مدل قرار است روابط بین لغات از نحوه قرارگیری آن‌ها در متون استخراج شود Word2vec .به دو صورت است. مدل کیسه لغات پیوسته[3] و اسکیپ گرام.[4]

از لحاظ الگوریتمی این دو روش شبیه هم هستند با این تفاوت که کیسه لغات پیوسته لغات هدف را از روی لغات متن ورودی پیش‌بینی می‌کند ولی اسکیپ گرام به صورت برعکس از روی لغات مرجوعه هدف، لغات ورودی را پیش‌بینی می‌کند.

[1] Knowledge Discovery and Data Mining
[2] Word embedding
[3] (CBOW)
[4] (Skip Gram)

برعکس کردن این چرخه دلخواه به نظر می‌رسد ولی از لحاظ آماری کیسه لغات پیوسته تأثیر نرمی بر روی همه اطلاعات توزیعی دارد (با رفتاری شبیه به یک مشاهده بر روی کل متن) و درکل این روش می‌تواند روشی مفید برای استفاده در مجموعه دادگان کوچک‌تر باشد. اما اسکیپ گرام با هر زوج محتوا–هدف به‌صورت یک مشاهده جدید رفتار می‌کند و در مجموعه دادگان بزرگ‌تر بهتر جواب می‌دهد.

2-3- پیشینه تحقیق

به‌طور کلی تاریخچه پردازش زبان طبیعی از دهه ۱۹۵۰ میلادی شروع می‌شود. در ۱۹۵۰ آلن تورینگ مقاله معروف خود را درباره آزمایش تورینگ که امروزه به عنوان ملاک هوشمندی شناخته می‌شود، منتشر ساخت . نخستین تلاش‌ها برای ترجمه توسط رایانه ناموفق بودند، به‌طوری‌که ناامیدی بنگاه‌های تأمین بودجه پژوهش از این حوزه را نیز در پی داشتند. پس از اولین تلاش‌ها آشکار شد که پیچیدگی زبان بسیار بیشتر از چیزی‌ست که پژوهشگران در ابتدا پنداشته‌بودند. بی‌گمان حوزه‌ای که پس از آن برای استعانت مورد توجه قرار گرفت زبان‌شناسی بود. اما در آن دوران نظریه‌ی زبان‌شناسی وجود نداشت که بتواند کمک شایانی به پردازش زبان‌ها بکند. در سال ۱۹۵۷ کتاب ساختارهای نحوی اثر نوام چامسکی زبان‌شناس جوان آمریکایی که از آن پس به شناخته‌شده‌ترین چهره زبان‌شناسی نظری تبدیل شد به چاپ رسید .از آن پس پردازش زبان با حرکت‌های تازه‌ای دنبال شد اما هرگز قادر به حل کلی مسئله نشد .

اخیرا داده کاوی موضوع بسیاری از مقالات ، کنفرانس ها و رساله ها ی عملی شده است ، اما این واژه تا اوایل دهه نود مفهومی نداشت و به کار برده نمی‌شد .در دهه شصت و پیش از آن زمینه‌هایی برای ایجاد سیستم ها ی جمع آوری و مدیریت داده‌ها ایجاد شد و تحقیقاتی در این زمینه انجام پذیرفت که منجر به معرفی و ایجاد سیستم های مدیریت پایگاه داده‌ها گردید.

ایجاد و توسعه مدلهای داده ای برای پایگاه سلسله مراتبی ، شبکه ای و بخصوص رابطه ای در دهه هفتاد ، منجر به معرفی مفاهیمی همچون شاخص گذاری و سازماندهی داده ها و در نهایت ایجاد زبان پرسش SQL در اوایل دهه هشتاد گردید تا کاربران بتوانند گزارشات و فرمهای اطلاعاتی مورد نظر خود را ، از این طریق ایجاد نمایند توسعه سیستم‌های پایگاهی پیشرفته در دهه هشتاد و ایجاد پایگاه‌های شی‌گرا، کاربردگرا و فعال باعث توسعه همه جانبه و کاربردی شدن این سیستم ها در سراسر جهان گردید . بدین ترتیب DBMS هایی همچون DB2، Oracle، Sybase و... ایجاد شدند و حجم زیادی از اطلاعات با استفاده از این سیستم ها مورد پردازش قرار گرفتند . شاید بتوان مهمترین جنبه در معرفی داده کاوی را مبحث کشف دانش از پایگاه داده ها (KDD) دانست بطوری که در بسیاری موارد DM و KDD بصورت مترادف مورد استفاده قرار می‌گیرند . برای اولین بار مفهوم داده‌کاوی در کارگاه IJCAI در زمینه KDD توسط Shapir مطرح گردید . به دنبال آن

در سالهای 1991 تا 1994 ، کارگاه‌های KDD مفاهیم جدیدی را در این شاخه از علم ارائه کردند بطوری که بسیاری از علوم و مفاهیم با آن مرتبط گردیدند.

برخی از کاربردهای داده کاوی در محیطهای واقعی عبارتند از :

1. خرده فروشی : از کاربردهای کلاسیک داده کاوی است که می‌توان به موارد زیر اشاره کرد :
   - تعیین الگوهای خرید مشتریان
   - تجزیه و تحلیل سبد خرید بازار
   - پیشگویی میزان خرید مشتریان از طریق پست(فروش الکترونیکی)
2. بانکداری :
   - پیش بینی الگوهای کلاهبرداری از طریق کارت‌های اعتباری
   - تشخیص مشتریان ثابت
   - تعیین میزان استفاده از کارت‌های اعتباری بر اساس گروه‌های اجتماعی
3. بیمه :
   - تجزیه و تحلیل دعاوی
   - پیشگویی میزان خرید بیمه نامه‌های جدید توسط مشتریان
4. پزشکی :
   - تعیین نوع رفتار با بیماران و پیشگویی میزان موفقیت اعمال جراحی
   - تعیین میزان موفقیت روشهای درمانی در برخورد با بیماری‌های سخت

امروزه کسب و کارهای بسیاری برای ارائه خدمات متنوع و تعامل با مشتری از رسانه‌های اجتماعی مانند فیس بوک و توئیتر استفاده می‌کنند. برای افزایش مزیت رقابتی و ارزیابی موثر محیط رقابتی کسب و کار، شرکت‌ها نیاز دارند که نه‌تنها ناظر مفاهیم ایجاد شده‌ی مشتریان در سایت‌های مربوط به رسانه‌های اجتماعی خود باشند، بلکه بر اطلاعات متنی ایجاد شده‌ی رقیبانشان در سایت‌های اجتماعی نیز نظارت کنند(هی،ژا و لی،2013)[1]. همانطور که اشاره شد شبکه‌های اجتماعی راه‌های جدیدی برای برقراری ارتباط میان افرادی با فرهنگ‌ها، ارزش‌های اجتماعی گوناگون فراهم کرده‌اند. این وبسایت‌ها ابزار بسیار قدرتمندی برای ایجاد ارتباط میان افراد و و به‌اشتراک گذاری دانش میان آنها هستند. در بسیاری از این شبکه‌های اجتماعی آنچه بیش از هر چیز به چشم می‌آید، استفاده از متن‌ها به عنوان ابزار انتقال دانش است. البته توجه به این نکته ضروری است که

[1] Hee Jae Lee

افراد در زندگی روزمره‌ی خود در شبکه‌های اجتماعی که که جزء جدایی ناپذیری از زندگی روزمره‌ی افراد هستند، به نحوه ی تلفظ و و نکات گرامری متن‌ها توجه زیادی نمی‌کنند. و همین مسئله استخراج الگوهای منطقی و اطلاعات دقیق از میان این متن‌های منتشر شده و به اصطلاح غیر ساخت یافته را با مشکل روبرو می‌کند. متن کاوی پاسخی به موضوعات ارائه شده در بالاست [8].

مطالعه‌ی جی گولبک [1]و همکاران در مقاله [9] اولین پژوهشی است که سعی دارد رفتارهای شخصیتی و پروفایل‌های شبکه‌های اجتماعی را به هم مربوط کند. ابتدا، نویسندگان یک فرم توییتر را تشکیل دادند که حاوی فهرست یک مدل شخصیتی به نام پنج بزرگ[2] شامل 45 سوال بود. برای هر شخص 2000 توئیت اخیر و فهرست شخصیتی پنج بزرگ آنها را برای مدل جمع آوری کرده‌اند. همچنین برای استخراج اطلاعات زبانی از پیام‌های آنها، از پایگاه داده روان شناختی (MRC) و تحقیق زبانی و تعداد کلمه (LIWC) استفاده کردند. سپس نتایج آزمون شخصیتی و نتایج استخراج اطلاعات زبانی به جدول همبستگی داده شد و بعد از آن برای شناسایی شخصیت از روشهای گاوسی و ZeroR استفاده شده است .

در مقاله [10] محققی به نام کوریکا[3] در سال 2011 تلاش دارد که امتیازهای شخصیتی را با کاربران توییتر پیوند دهد. آنها این کار را با جمع‌آوری داده از یک اپلیکیشن فیسبوک به نام myPersonality انجام دادند. حدود 40٪ کاربران این اپلیکیشن اجازه دادند تا پروفایلها و امتیازهای شخصیتیشان به اشتراک گذاشته شود. سپس نویسندگان تنها کسانی را که حساب توییتر خود را در نمایه فیس بوک خود مشخص کرده اند، در مجموعه داده در نظر گرفتند. سپس آزمون شخصیتی Big-Five پنج بزرگ بر روی آن افراد انجام شد. آنها ارتباط میان صفات شخصیت Big-Five و پنج نوع از کاربران میکروبلاگ را تجزیه و تحلیل کردند: کاربران محبوب، بیشتر خوانده شده، شنوندگان؛ و دو عامل تاثیرگذار یعنی زمان و Klout. با استفاده از این، نویسندگان یک جدول همبستگی ایجاد کردند و سپس با استفاده از الگوریتم M5 Rules، رگرسیون را برای پیش بینی شخصیت پروفایل‌ها انجام دادند.

در مطالعه [11]، برای شناسایی شخصیت، نویسندگان ویژگی‌های متنی و جمعیت‌شناسی را از پروفایل های فیس بوک استخراج کردند. برای پیش بینی شخصیت بر اساس شاخص شخصیت پنج عامل شخصیتی، هر یک از کاربران میبایست به 45 سوال پاسخ بدهند. سپس صفاتی مانند جنسیت، سن، محل، نقل قول ها، وضعیت ارتباط، عکس ها، نظرات و غیره به منظور تعریف هر فرد استفاده شدند. سپس با استفاده از این اطلاعات، افراد با توجه به پنج فاکتور شخصیتی رتبه بندی شدند، بر اساس اینکه افراد هر صفت شخصیتی را بالای 5

[1] J. Golbeck
[2] Big Five
[3] Quercia, M

الی 10 درصد دارند یا خیر. برای این کار آنها مدل‌های پیش بینی عددی شامل جداول تصمیم‌گیری، رگرسیون-خطی و REPTree را به‌کاربردند.

در مقاله [12]، نویسندگان معماری جدیدی را پیشنهاد کردند تا شخصیت را با استفاده از مفاهیم ساده همراه با برچسب‌های عاطفی مرتبط و تضاد عاطفی، شناسایی کنند. آنها از مجموعه مقالات حاوی 2400 مقاله که به صورت دستی برای مدل پنج فاکتور شخصیتی برچسب گذاری شده اند استفاده می کنند. همچنین نویسندگان ویژگی‌های استخراج شده از متن توسط LIWC، MRC را با ویژگی‌های مبتنی بر دانش عامیانه استخراج شده توسط تکنیک‌های محاسباتی سنتی ترکیب می کنند.

مقاله[13] نقاط ضعف روش یادگیری نظارت‌شده را برجسته می‌کند. این تحقیق نشان می‌دهد که مشکل یادگیری نظارت شده دسترسی محدود و هزینه‌ی زیاد برای پیدا کردن داده‌های برچسب‌دار شده برای آموزش است. بنابراین این مطالعه روشی را پیشنهاد می کند که مربوط به یادگیری گروهی است. در اینجا، طبقه بندی‌ها با استفاده از اطلاعات استخراج شده از مجموعه داده‌های مختلف که از ژانرهای مختلف، چندین زبان و سیستم های مختلف پیش بینی شخصیت هستند، ساخته شده‌اند.

در مقاله[14] که توسط دانشگاه استنفورد انجام شد برای تشخیص شخصیت بر اساس مدل شخصیتی MBTI روشی جدید ارائه گردید که با دقت مناسی توانسته هر یک از 16 کلاس‌های شخصیتی MBTI را با دقتی بالای 37 درصد تشخیص دهد.

2-4- جمع بندی فصل: چارچوب نظری پژوهش

بر خلاف تصور عمومی، نظریه و تحقیق به عرصه هایی مجزا تعلق ندارند، بلکه مکمل یکدیگرند . نظریه القاء کننده فرضیه‌ها و در خلال حل یک مساله نظری می تواند موجد افکار بیشتری شود(نائینی،1367).

شخصیت را می‌توان آن الگوهای معین و مشخصی از تفکر، هیجان و رفتار تعریف کرد که سبک شخصی فرد را در تعامل با محیط اجتماعی و مادی‌اش رقم می‌زنند. به عبارت دیگر شخصیت شامل ویژگی‌های نسبتاً با ثبات و پایدار است که در توصیف آن‌ها از صفاتی همچون زودرنج، مضطرب، پر حرف، درونگرا و برون‌گرا و غیره استفاده می‌شود.(اتکینسون و هیلگارد،1953)

شخصیت عامل اصلی نحوه‌ی رفتار افراد در اینترنت است. از آنجا که شبکه به دلیل ماهیت خود متشکل از تعاملات انسانی است، نتیجه میشود که ما نمی توانیم نحوه ی عملکرد اینترنت را بدون درک شخصیت افرادی که در آن تعامل میکنند بفهمیم.

درک شخصیت افراد توسط کامپیوتر نیاز به مدلی دارد که در آن کلمات و جملات مفهوم داشته باشند و سپس کامپیوتر با روش‌های یادگیری ماشین بتواند این مدل را حین فرایند یاد بگیرد. ما از الگوریتم‌های یادگیری نظارت شده برای ارائه‌ی بهترین مدل استفاده خواهیم کرد.

ورودی این الگوریتم توئیت‌ها و یا هر پستی که افراد به صورت متنی در فضای مجازی منتشر کرده‌اند بعلاوه تیپ شخصیتی آن‌ها می‌باشد و مدل بایستی این ورودی‌ها را یاد بگیرد.(آمیچی-هامبورگر،2002)[1]

امروزه رشد و تغییر نمایی استفاده از شبکه‌های اجتماعی مجازی موجب شده‌است که افراد روزانه اطلاعات زیادی به اشتراک بگذارند. این اطلاعات شامل تصویر، صدا، متن و غیره می‌باشد. شبکه‌های اجتماعی یکی از بهترین منابع برای تحلیل و آنالیز پدیده‌های مختلف محسوب میشوند. یک چالش بحث برانگیز محققین امروز این است که آیا این داده‌ها میتوانند حاوی اطلاعاتی در مورد شخصی که آنها را به اشتراک گذاشته‌است باشند؟ الگوهای رفتاری‌ای که اشخاص از آنها پیروی میکنند ناشی از شخصیت درونشان است. این الگوهای رفتاری شامل حرف زدن و نوشتار هم می‌شود. به این معنی که اشخاص با شخصیت‌های مختلف نوشتار مختلفی دارند. ما برای این کار ابتدا باید مدل‌های شخصیتی استاندارد را بشناسیم. یکی از معتبرترین مدل‌ها مدل MBTI است. پدیدآورندگان این مدل کاترین کوک بریگز و دخترش، ایزابل بریگز مایرز بودند که بر اساس تحقیقاتشان روی مطالعات یونگ این آزمون را طراحی کردند. آن‌ها این آزمون را ابتدا در طی جنگ جهانی دوم بدین منظور ارائه کردند تا بتوانند مناسب‌ترین شغل را برای زنانی که در صنعت نظامی کار می‌کردند پیدا کنند. بر اساس این مدل افراد در 16 کلاس شخصیتی تقسیم می‌شوند. این مدل برای شخصیت 4 فاکتور را در نظر میگیرد که عبارتند از:

1-درونگرا- برونگرا

2-شمی – حسی

3-منطقی-احساسی

4-قضاوتی- ادراکی

هر فرد در آن واحد می‌تواند از هر فاکتور شخصیتی در یک گروه قرار بگیرد. برای مثال یک فرد می‌تواند درونگرا، شمی، احساسی، قضاوتی باشد.

16 کلاس شخصیتی این مدل عباتند از:

INTP-INTJ-INFP-INFJ-ISTP-ISTJ-ISFP-ISFJ-ENTP-ENTJ-ENFP-ENFJ-ESTP-ESTJ-ESFP-ESFJ

این فاکتورها در نوشتار تاثیر میگذارند. هدف این پایان‌نامه یافتن مدل و الگوریتمی است که بتواند بر اساس نوشتار افراد در شبکه‌های اجتماعی فاکتورهای شخصیتی آنان را حدس بزند. بنابراین ابتدا باید کلمات را برای کامپیوتر طوری تعریف کنیم که معنی آنها را نیز در بر داشته باشد. تا با استفاده از آن بتواند جملات را تفسیر و کلاس‌بندی کند.

[1] Amichai-Hamburger

فرآیند کلی ساختن مدل شامل سه مرحله است:

1.آماده سازی داده‌ها(پیش‌پردازش داده)

2.تفسیر داده‌ها برای کامپیوتر

3.ساخت و ارزیابی مدل

در فاز اول داده‌ها را در اصطلاح باید تمیز کنیم. نویز ها برطرف و اطلاعات غیر مفید شامل آدرس وبسایت هایی که پست افراد در آنها بوده‌است حذف باید بشوند. همچنین توزیع داده‌ها را در کلاسها برای آموزش یکسان باید باشند . و در نهایت داده برای ورود به سیستم آماده می‌شوند.

در مرحله‌ی دوم فایل داده‌ای که پیش پردازش شده است باید خوانده شود و کلمات آن را با بردارهای عددی برای مدل نمایش دهیم هر کلمه را به صورت بردار به قسمی نمایش میدهیم که کلماتی که هم معنی هستند و یا معانی نزدیک تر و مرتبط تری به هم دارند بردارهایشان به هم نزدیکتر باشند. همچنین کلماتی که بیشتر با هم ممکن است در یک متن ظاهر شوند نیز بردارهای نزدیک به همی داشته باشند.

در مرحله‌ی سوم مدل باید طراحی شود و بردارهای کلماتی را که آماده شده‌اند برای کلاس‌بندی به مدل تزریق می‌شوند تا الگوی مورد نظر را برای این مسئله با کمترین میزان خطای ممکن بدست بیاید.

فصل سوم:

روش اجرایی تحقیق

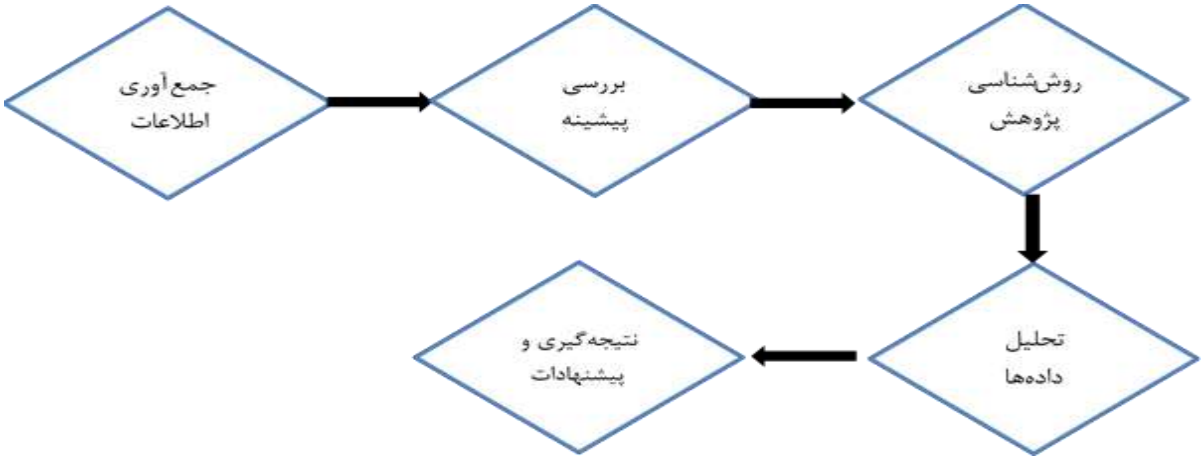
3-1- مقدمه

این بخش به بررسی روش مطالعه و پژوهش در پایان نامه با رویکرد شبکه عصبی می‌پردازد. در این فصل تلاش شده چارچوب کلی روش پژوهش در این پایان نامه ارائه گردد و همچنین در این فصل، روش‌شناسی پژوهش از نقطه نظر هدف پژوهش، نوع پژوهش، زمان، جامعه و نمونه آماری و روش نمونه‌گیری، روش گردآوری اطلاعات، مراحل انجام پایان‌نامه، شیوه دستیابی به متغیرها و مفاهیم مدل مفهومی پژوهش، روایی و پایایی ابزارهای گردآوری اطلاعات و... ، به‌صورت مشروح توضیح داده می‌شوند.

3-2- روش تحقیق

روش انجام این پایان‌نامه از نظر هدف، کاربردی است، و از نظر ماهیت و روش، تجربی است. زیرا هدف این نوع تحقیق ها بررسی تأثیر محرک‌ها، روش‌ها و یا شرایط خاص محیطی بر روی یک گروه آزمودنی می‌باشد. ما نیز در تحقیق خود می‌خواهیم تاثیر شخصیت افراد را بر روی نوشتار آنها را بررسی کنیم. مراحل انجام پژوهش حاضر، به شرح ذیل می‌باشد:

1- جمع‌آوری‌اطلاعات
2- بررسی پیشینه ی تحقیق و تحقیق‌های مرتبط
3- انتخاب روش و مدل تحقیق
4- اجرای مدل و آنالیز داده ها
5- تحلیل نتایج و پیشنهادات



شکل 3-1- فرآیند انجام پژوهش

3-3- متغیرهای تحقیق

متغییرهای وابسته:

در بین کلماتی که افراد استفاده می‌کنند روابطی وجود دارد که در نتیجه استفاده از آنها مستقل از هم نیستند.

متغیرهای مستقل:

در این تحقیق ما کلماتی که افراد برای نوشتار استفاده می‌کنند را در نظر گرفته‌ایم.

3-4- جامعه آماری

از آنجایی که این پژوهش موضوعی بسیار نو و به روزی دارد جمع آوری داده برای آن کار مشکلی می‌باشد. برای انجام این پژوهش جامعه‌ی آماری مورد مطالعه‌ی ما افرادی هستند که به صورت تصادفی و بدون هیچ فرضی انتخاب شده‌اند. این افراد در سایت www.personalitycafe.com تست‌های شخصیتی را انجام داده‌اند و به مدیران سایت اجازه‌ی استفاده از داده‌های پروفایل‌های آنان در شبکه های مجازی مختلف مانند توئیتر، فیس بوک، یوتیوب و غیره را داده‌اند.

3-5- روش نمونه گیری

روش نمونه گیری به صورت تصادفی از اطلاعات کاربران در فضای مجازی می‌باشد که این داده در سایت www.kaggle.com برای یک مسابقه آپلود شده‌است.

مجموعه داده ما شامل 8600 نفر می‌باشد که برای هر فرد 50 پست متنی در اختیار داریم. ما داده را به نسبت 80 به 20 برای آموزش و تست مدل تقسیم کرده ایم. با 7000 داده مدل را آموزش می‌دهیم و با باقی مانده آن‌را ارزیابی می‌کنیم. برای بهتر شدن مدل تمام داده‌های آموزشی را به دسته های کوچکتر تقسیم می‌کنیم و در هر مرحله‌ی آموزش قسمت کوچکی از داده را برای آموزش به سیستم می‌دهیم. در هر مرحله وزن ها تغییر می‌کنند تا این فرایند همگرا شود. سپس هنگامی که تمامی قسمت‌های مجموعه داده آموزشی را یک بار به مدل دادیم و آن را ارزیابی کردیم می‌توانیم این کار آنقدر تکرار کنیم تا درصد دقت بالاتر رود. تعداد قسمت‌های داده آموزش و همچنین تعداد تکرارهای الگوریتم ازجمله فرا پارامترهایی هستند که م با آنها مواجه هستیم و باید آنها را بسته به شرایط مسئله آنقدر تغییر دهیم تا به مقدار بهینه برای مسئله برسیم.

3-6- قلمرو تحقیق

قلمروهای پژوهش حاضر عبارت است از :

3-6-1- قلمرو زمانی

این مجموعه داده در سال 2017 در سایت www.kaggle.com آپلود شده‌است.

3-6-2- قلمرو موضوعی

قلمرو موضوعی این پژوهش در زمینه ″ ارائه مدلی جهت پیش‌بینی شخصبت افراد بر اساس متن منتشر شده از شبکه‌های اجتماعی.″ با استفاده از مدل سازی شبکه عصبی مصنوعی و یادگیری نظارت شده می‌باشد.

3-7- روش گردآوری داده‌ها

در این تحقیق اطلاعات اولیه که به صورت داده‌ی متنی شامل تیپ شخصیتی افراد و 50 عدد از پست‌هایشان به زبان انگلیسی می‌باشد از سایت داده کاوی www.kaggle.com گرفته شده است. از این افراد اطلاعات دیگری در مجموعه داده وجود ندارد.

برای مدل MBTI این مجموعه داده بزرگترین مجموعه داده ی موجود است. ما برای هر فرد 50 پست از او را داریم هدف کلاس بندی این متون است بگونه ای که برچسب کلاس با برچسبی که برای شخصیت فرد در نظر گرفته شده است برابر باشد.

3-8- ابزار جمع‌آوری داده‌ها

جمع آوری داده در سایت www.personalitycafe.com که یک سایت تحلیل شخصیتی است و انواع مختلف تست‌های شخصیتی را دارد به صورت یک پرسشنامه اینترنتی بوده‌است که در آن تیپ شخصیتی افراد بر اساس آزمون MBTI مشخص می‌شود. سپس اطلاعات کاربری افراد در شبکه‌های مجازی از آنها گرفته می‌شود و از هر فرد آخرین پست‌های متنی آنها در شبکه های مجازی به همراه آدرس اینترنتی آن و نوع تیپ شخصیتی فرد در یک فایل ذخیره می‌شوند.

ما برای بسیاری از کارهای پردازش متن وNLP ، نیاز به نمایش عددی کلمات و متون داریم تا بتوانیم از انواع روش‌های عددی حوزه یادگیری ماشین مانند اکثر الگوریتم های دسته بندی روی لغات و اسناد استفاده کنیم. یکی از رهیافت هایی که در این حوزه بسیار رایج است ، نمایش برداری کلمات و جملات است. روشی که

توسط گوگل در سال ۲۰۱۳ پیشنهاد شده است و روشی بسیار کارآمد و مناسب برای نمایش لغات و متون و پردازش آنها است روش Word2Vec است که ما در پژوهش خود از آن استفاده کرده‌ایم.
در این روش به کمک شبکه عصبی یک بردار با اندازه کوچک و ثابت برای نمایش تمام لغات و متون در نظرگرفته‌شده و با اعداد مناسب در فاز آموزش مدل یا training برای هر لغت این بردار محاسبه می‌شود. در این بردار هر عدد ، نمایشگر ویژگی خاصی نیست و فقط یک عدد را نمایش می دهد.
پس از گذشتن از این مرحله ما داده‌های خود را که کلمات و جملات هستند به صورت بردارهای عددی درآورده‌ایم و برای هر شخص چندین بردار که نمایش دهنده پست آن شخص است، خواهیم داشت.

3-9- روش تجزیه و تحلیل داده‌ها

ابزار تجزیه و تحلیل داده‌ها پایان‌نامه مذکور شبکه‌های عصبی مصنوعی هستند. یک شبکه عصبی مصنوعی ابزاری است برای مدل سازی و تحلیل داده‌ها با الهام از شبکه عصبی طبیعی. با در نظر گرفتن این حقیقت مدل های ریاضی مختلفی که توصیف‌گر این پردازش باشند، ارائه شده است .در واقع، شبکه‌های عصبی مصنوعی را می‌توان به صورت یک مدل ریاضی یا سیستمی که شامل چندین المان پردازشی ساده به نام نرون که به صورت موازی در یک یا چند لایه با معماری‌های گوناگون عمل می‌کنند، تعریف کرد. بنابراین شبکه عصبی روشی برای محاسبه است که بر پایه اتصال به هم پیوسته چندین واحد پردازشی ساخته می‌شود.
شبکه‌های عصبی کانولوشن یکی از مهمترین روش های یادگیری عمیق هستند که در آنها چندین لایه با روشی قدرتمند آموزش می‌بینند. این روش بسیار کارآمد بوده و یکی از رایج‌ترین روش‌ها در کاربردهای مختلف است. بطور کلی، یک شبکه کانولوشن از سه لایه اصلی تشکیل میشود که عبارتند از : لایه کانولوشن، لایه Pooling و لایه تماما متصل. لایه‌های مختلف وظایف مختلفی را انجام می‌دهند.
در هر شبکه کانولوشنی دو مرحله برای آموزش وجود دارد مرحله feedforward و مرحله backpropagation یا پس‌انتشار .در مرحله اول داده‌ی ورودی به شبکه تغذیه می‌شود و این عمل چیزی جز ضرب نقطه ای بین ورودی و پارامترهای هر نورون و نهایتا اعمال عملیات کانولوشن در هر لایه نیست. سپس خروجی شبکه محاسبه می‌شود. در این جا به منظور تنظیم پارامترهای شبکه و یا به عبارت دیگر همان آموزش شبکه، از نتیجه خروجی جهت محاسبه میزان خطای شبکه استفاده می‌شود. برای اینکار خروجی شبکه را با استفاده از یک تابع خطا با پاسخ صحیح مقایسه کرده و اینطور میزان خطا محاسبه می‌شود. در مرحله بعدی بر اساس میزان خطای محاسبه شده مرحله پس انتشار آغاز می‌شود. در این مرحله گرادیانت هر پارامتر با توجه به قائده زنجیره‌ای محاسبه می‌شود و تمامی پارامترها با توجه به تاثیری که بر خطای ایجاد شده در شبکه دارند تغییر

پیدا می‌کنند. بعد از بروزآوری شدن پارامترها مرحله بعدی feed-forward شروع می‌شود. بعد از تکرار تعداد مناسبی از این مراحل آموزش شبکه پایان می‌یابد.

شبکه‌های عصبی پیچشی یا کانولوشن نسبت به بقیه رویکردهای دسته‌بندی تصاویر به میزان کمتری از پیش‌پردازش استفاده می‌کنند. این امر به معنی آن است که شبکه معیارهایی را فرامی‌گیرد که در رویکردهای قبلی به صورت دستی فراگرفته می‌شدند. این استقلال از دانش پیشین و دستکاری‌های انسانی در شبکه‌های عصبی پیچشی یک مزیت اساسی است. توده خروجی شبکه‌های کانولوشن را می‌توان بصورت یک توده سه بعدی از نورون ها تفسیر کرد. به زبان ساده تر یعنی اینکه خروجی این لایه یک توده سه بعدی است.

مزیت این نوع تحلیل داده این است که ما دیگر نیاز به تعریف ویژگی‌های داده نداریم و داده را بصورت خام به مدل می‌دهیم و خود مدل سعی می‌کند تا ویژگی‌های خاص داده را کشف کند که گاهی این ویژگی ها برای ما غیر قابل درک هستند. در این نوع شبکه ها ما پس از هر لایه نیاز به یک تابع فعالسازی داریم که بر روی مقادیر محاسبه شده در نورون اعمال می‌شود و خروجی را به عنوان ورودی به لایه‌ی بعد می‌دهد. انتخاب این تابع فعالسازی بصورت تجربی است و بسته به نوع مسئله و شرایط حاکم بر آن می‌تواند متفاوت باشد. فاکتورهای زیادی روی انتخاب آن تاثیر گذار هستند. اما با دید مهندسی گاهی هم بعضی توابع فقط در عمل بهتر جواب می‌دهند.

در انتهای آموزش، شبکه به ما مدلی خواهد داد که در آن وزن‌ها آموزش دیده شده‌اند تا بیشترین درصد دقت ممکن را داشته باشند. این نوع شبکه ها بسیار مستعد هستند که داده‌های ورودی را حفظ کنند. و overfit رخ دهد. برای جلوگیری از این امر تدابیری در نظر گرفته شده‌است. که شامل: dropout، cross-validation، shrinkage و غیره می‌باشد.

3-10- خلاصه فصل

در این فصل به بیان روش پژوهش و روش تجزیه و تحلیل این پژوهش پرداخته و همچنین مراحل انجام این پژوهش به صورت مفصل بیان شد. در ادامه و در فصل چهارم مدل تعریف شده، و به ارائه مدلی جهت پیش‌بینی نوع شخصیتی افراد با استفاده از متن های منتشر شده از آنها در فضای مجازی استفاده از مدل سازی شبکه عصبی اقدام می‌گردد و در فصل پنجم نتایج و تحلیل نهایی ارائه خواهد شد.

فصل چهارم :

یافته‌های پژوهش

4-1- مقدمه

این بخش به توضیح و تشریح روش پیشنهادی برای انجام تحقیق و یافته‌های پژوهش اختصاص یافته‌است. همچنین تلاش شده در این بخش به تمام جزییات پیاده‌سازی و چالش های روبرو پرداخته‌شود. در ابتدا متغیرهای تحقیق را بیان می‌کنیم و سپس شیوه‌ی تبدیل داده‌ی خام به اطلاعات و سپس دانش را در پژوهش خود بیان خواهیم کرد. که شامل پیش‌پردازش داده‌ها، یکسان کرده توزیع داده‌ها و غیره می‌باشد. در ادامه مدل ارائه شده را به جزئیات ارائه خواهیم کرد و سپس نتایج بدست آمده را گزارش خواهیم کرد.

4-2- فرضیات تحقیق

همانطور که در فصل اول اشاره کردیم، فرض می‌کنیم الگوریتمی وجود دارد که شخصیت افراد را از روی گفتارشان تشخیص می‌دهد. به عبارتی دیگر فرض می‌کنیم که رابطه‌ی معناداری بین نوشتار افراد و شخصیتشان وجود دارد. به شیوه‌ای که افراد کلمات را به کار می برند مقدار زیادی اطلاعات از فرایندهای روان شناختی پایه شامل سرنخ هایی از افکار، احساسات، ادراکات و شخصیت شان را منعکس می‌سازند. کلمات می‌تواند نشانگر پایگاه اجتماعی، سن، انگیزه‌ها و گرایشات روان شناختی مثل خُلق افراد باشد. پژوهش های متعددی در این زمینه انجام شده‌اند که این فرضیه را به اثبات برسانند. بر اساس مقاله [15] میهل و نیدرهافر و پنه بکر(2003) گروم و پنه‌بکر(2002) بر این باورند که الگوهای کاربرد کلمه می توانند مثل یک اثرانگشت یا یک نمونهٔ DNA تیپ شخصیتی و هویت ویژه و خاص او را تعیین کنند.

بنابراین اگر نوشتار افراد شخصیتشان را نشان می‌دهد، نوشته‌های آنها در فضای مجازی هم می‌تواند شامل شخصیتشان شود، زیرا امروزه بخش عمده‌ی ارتباطات از طریق شبکه‌های اجتماعی صورت می‌گیرند.

همچنین فرض کردیم برای اینکه پژوهش معنادار باشد فرض کردیم که رابطه‌ای بین اطلاعات نوشتار افراد و نوع شخصیت آنها وجود دارد. که در این زمینه نیز پژوهش‌های زیادی صورت گرفته است که وجود این ارتباط را اثبات می‌کند. برای نمونه ابرلند و گیل (2006) و دوله و فارنهایم(1999) نشان داده‌اند که برونگراها از لحاظ لغوی غنا کمتری داشته، بیشتر صحبت می کنند و خطاها معناشناختی بیشتری در مقایسه با درون گراها دارند.

همچنین استخراج اطلاعات افراد از روی نوشتار توسط ماشین امکان‌پذیر است مطالعات زیادی در این زمینه انجام شده است و الگوریتم‌هایی برای این منظور با دقت تقریبا مناسبی ارائه شده‌است که براساس چند مدل شخصیتی که غالبا مدل پنج عامل شخصیتی هستند شخصیت را پیش‌بینی می‌کنند. این مطالعه هم همچنین به ارائه‌ی مدلی برای تشخیص مدل شخصیتی MBTI می‌پردازد.

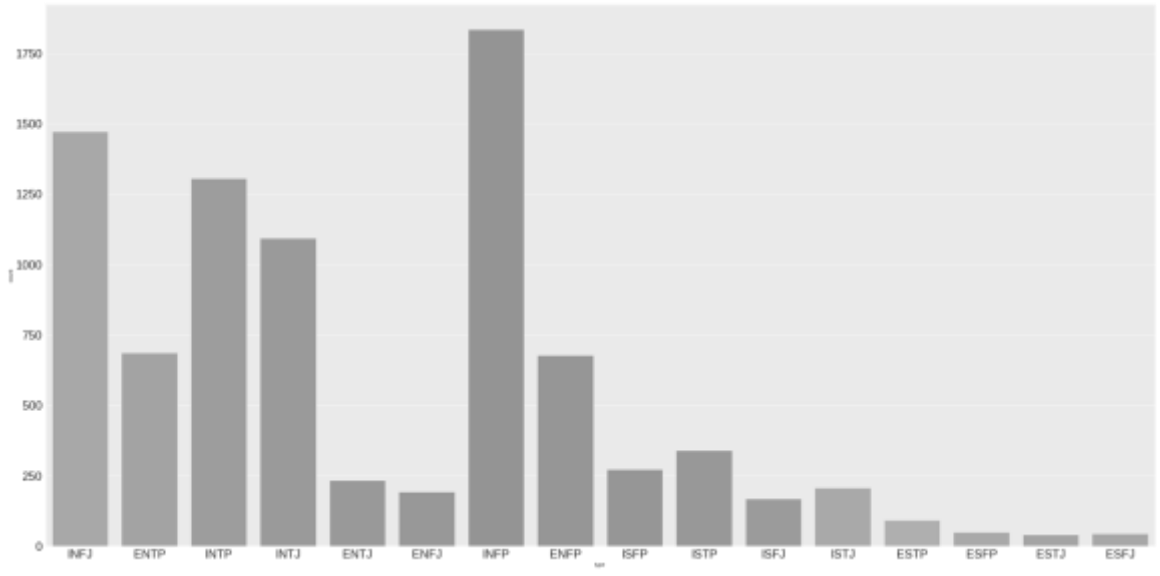
4-3- پیش‌پردازش داده‌ها

4-3-1- یکسان سازی توزیع داده‌ها

برای عمل کلاس‌بندی، داده‌های آموزشی باید در کلاس‌ها بصورت یکسان توزیع شده‌باشند. اگر توزیع داده‌ها برای آموزش یکسان نباشد، سیستم به سمت کلاسی که بیشترین تعداد داده را دارد می‌رود و وزن‌های سیستم متناسب با داده‌های این کلاس تنظیم می‌شود. زیرا هر داده‌ای که به سیستم داده می‌شود برای آموزش، وزن ها متناسب با آن داده به‌روزرسانی می‌شوند. بنابراین اگر تعداد داده‌های از یک نوع بیشتر از انواع دیگر باشد وزن ها متناسب با آن کلاس آموزش داده خواهند شد.

در برخی جوامع داده های موجود در کلاس ها به صورت طبیعی ممکن است توزیع یکسانی نداشته باشند و این خاصیت مسئله است، در این مواقع باید توزیع‌ها را به روش هایی که موجود هستند یکسان کرد. در داده‌ی مورد مطالعه در این پژوهش نیز توزیع داده‌ها یکسان نیستند. در شکل زیر توزیع داده‌ها در کلاس‌های مختلف نشان داده شده‌است.

درمیان این روش های بازنمونه‌گیری، ما از روش downsampling استفاده کردیم به این صورت که تعداد داه‌ها را در همه‌ی کلاس‌ها محاسبه کردیم . در شکل 4-1 توزیع داده‌ها در کلاس‌های مختلف نشان داده شده است. همانطور که در شکل مشخص است داده‌ها توزیع بسیار متفاوتی دارند بنابراین نمی‌توان بدون یکسان‌سازی توزیع‌ها از این داده استفاده کرد.



شکل 4-1 نحوه توزیع پست‌ها در کلاس های شخصیتی 16 گانه

برای اینکه اطلاعات کمتری را از دست دهیم، ما توزیع داده‌ها به این صورت تنظیم کردیم که در هر بعد شخصیت به‌صورت مساوی تقسیم شده‌باشند.
در مجموعه داده ما برای هر فرد 50 پست در اختیار داریم. ما این پست‌ها را از هم جدا کردیم و در واقع 50×7600 پست و نوع تیپ شخصیتی برای آموزش داریم و قرار بر این است که از یک پست متنی شخصیت پیش‌بینی شود.

4-3-2- بازنمایی کلمات

برای اینکه بتوانیم داده‌ها را به عنوان ورودی به شبکه بدهیم نیاز داریم که آنها را تعریف کنیم. بهترین راه برای تعریف داده‌های متنی تبدیل آنها به اعداد است.
برای تبدیل کلمات به اعداد چند راه داریم. اول اینکه یک دیکشنری از تمام کلمات موجود در داده بسازیم، و به هر کلمه یک مقدار حقیقی به صورت تصادفی بدهیم. روش دیگراین است که کلمات را به صورت بردارها نمایش دهیم. در این روش نیز می‌توانیم یک دیکشنری بسازیم و به هر کلمه یک بردار نسبت بدهیم که این بردارها از یک توزیع گاوسی با میانگین و انحراف معیار معینی تولید می‌شوند. زیرا اگر بردار همه‌ی کلمات در یک بازه نباشند و به هم نزدیک نباشند فرایند آموزش دچار ایراد می‌شود.
مشکلی که این روش‌ها دارند این است که اعدادی که برای نمایش کلمات انتخاب می‌شوند هیچ اطلاعاتی را در باره‌ی آن نمی‌دهند زیرا کاملا تصادفی انتخاب شده‌اند. بنابراین با داشتن بردار یک کلمه هیچ اطلاعاتی از آن نخواهیم داشت و تفاوت بردارهای کلمات، تفاوت معنی آنها را نشان نمی‌دهد. برای اینکه بردارهای عددی نماینده‌ی کلمات برای کامپیوتر معنا داشته باشند باید روابط بین بردارها معنی‌دار باشد. برای این کار الگوریتم-هایی وجود دارد که ما از آنها استفاده کرده‌ایم.

4-3-2-1- الگوریتم بازنمایی کلمات

4-3-2-1-1- ساخت دیکشنری

برای بسیاری از روشهای پردازش متن وNLP ، نیاز به نمایش عددی کلمات و متون داریم تا بتوانیم از انواع روشهای عددی حوزه یادگیری ماشین مانند اکثر الگوریتم های دسته بندی روی لغات و اسناد استفاده کنیم. یکی از رهیافت هایی که در این حوزه بسیار رایج است ، نمایش برداری کلمات و جملات است.
فرض کنید فرهنگ لغتی داریم با N کلمه و لغت که به ترتیب الفبایی مرتب شده اند و هر لغت یک مکان مشخص در این فرهنگ لغت دارد. حال برای نمایش هر کلمه، برداری در نظر میگیریم با طول N که هر خانه

آن، متناظر با یک لغت در فرهنگ لغت ماست که برای راحتی کار فرض می کنیم شماره آن خانه بردار، همان اندیس لغت مربوطه در این فرهنگ لغت خواهد بود . با این پیش فرض، برای هر لغت ما یک بردار به طول Nداریم که همه خانه های آن بجز خانه متناظر با آن لغت صفر خواهد بود. در خود ستون متناظر با لغت عدد یک ذخیره خواهد شد (One-Hot encoding) با این رهیافت، هر متن یا سند را هم می‌توان با یک بردار نشان داد که به ازای هر کلمه و لغتی که در آن به کار رفته‌است ، ستون مربوطه از این بردار برابر تعداد تکرار آن لغت خواهد بود و تمام ستون های دیگر که نمایانگر لغاتی از فرهنگ لغت هستند که در این متن به کار نرفته‌اند، برابر صفر خواهد بود.

به این روش نمایش متون، صندوقچه کلمات یا Bag Of Words می‌گویند که بیانگر این است که برای هر لغت در صندوقچه یا بردار ما، مکانی درنظر گرفته شده است.

با این روش ما دو بردار عددی داریم که حال می توانیم از این دو در الگوریتم های عددی خود استفاده کنیم. با وجود سادگی این روش ، اما معایب بزرگی این بر آن مترتب است. مثلاً اگر فرهنگ لغت ما صد هزار لغت داشته باشد ، به ازای هر متن ما باید برداری صد هزارتایی ذخیره کنیم که هم نیاز به فضای ذخیره سازی زیادی خواهیم داشت و هم پیچیدگی الگوریتم ها و زمان اجرای آنها را بسیار بالا می برد.

از طرف دیگر در این نحوه مدلسازی ما فقط کلمات و تکرار آنها برای ما مهم بوده است و ترتیب کلمات یا زمینه متن (اقتصادی ، علمی ، سیاسی و ...) تاثیری در مدل ما نخواهد داشت.

4-3-2-1-2- روش word2vec

روشی دیگر که توسط گوگل در سال ۲۰۱۳ پیشنهاد شده است و روشی بسیار کارآمد و مناسب برای نمایش لغات و متون و پردازش آنها است روش Word2Vec است.در این روش به کمک شبکه عصبی یک بردار با اندازه کوچک و ثابت برای نمایش تمام لغات و متون در نظرگرفته شده و با اعداد مناسب در فاز آموزش مدل یا training برای هر لغت این بردار محاسبه می شود. در این بردار هر ستون، نمایشگر کلمه یا ویژگی خاصی نیست و فقط یک عدد را نمایش می‌دهد. اگر این بردار را ۴۰۰ تایی فرض کنید، یک فضای ۴۰۰ بعدی خواهیم داشت که هر لغت در این فضا یک نمایش منحصر بفرد خواهد داشت. برای افزایش دقت این روش، مجموعه داده اولیه که برای آموزش مدل مورد نیاز است، باید حدود چند میلیارد لغت را که درون چندین میلیون سند یا متن به کار رفته‌اند، در برگیرد. بعد از ایجاد بردارهای مرتبط با هر لغت، برای نمایش برداری هر متن یا خبر، می‌توان بردار تک تک کلمات به کار رفته در آنرا یافته و میانگین اعداد هر ستون را به دست آورد که نتیجه آن یک بردار برای هر متن یا سند خواهد بود.

سرعت این آموزش هم بسیار بالاست و در عرض چند ساعت و یا چند دقیقه (بسته به این که از کدام یک از دو الگوریتم آموزش آن استفاده کنیم) می‌توان حجم عظیمی از داده ها را به این الگوریتم داد و بردارهای لغات را ایجاد کرد .

این روش که الگوریتم آن به صورت متن باز نیز منتشر شده است و کتابخانه‌های مختلفی برای زبان‌های مختلف برای کار با آن تولید شده‌است، زمانی که توسط گوگل بر روی حجم بالای متون و اطلاعات به کار رفته است ، نتایج بسیار شگرفی را به همراه داده است.

مثلا اگر بردار لغت پادشاه را منهای بردار لغت مرد کنیم ، نتیجه به بردار کلمه ملکه بسیار نزدیک است.

در ادامه توضیح خواهیم داد که این روش چگونه کار می‌کند.

Skip-gram و continuous bag-of-words (CBOW) دو روش برای این الگوریتم هستند. این دو روش که هر دو یک شبکه عصبی ساده هستند که بدون وجود لایه پنهانی که در اغلب روش‌های شبکه عصبی وجود دارد، به کمک چند قانون ساده، بردارهای مورد نیاز را تولید می‌کنند. در روش کیف لغات پیوسته(CBOW) ، ابتدا به ازای هر لغت یک بردار با طول مشخص و با اعداد تصادفی (بین صفر و یک) تولید می‌شود. سپس به ازای هر کلمه از یک سند یا متن، تعدادی مشخص از کلمات بعد و قبل آنرا به شبکه عصبی می‌دهیم (به غیر از خود لغت فعلی) و با عملیات ساده ریاضی، بردار لغت فعلی را تولید می کنیم (یا به عبارتی از روی کلمات قبل و بعد یک لغت، آنرا حدس می زنیم) که این اعداد با مقادیر قبلی بردار لغت جایگزین می‌شوند. زمانی که این کار بر روی تمام لغات در تمام متون انجام گیرد، بردارهای نهایی لغات همان بردارهای مطلوب ما هستند.

روش Skip-gram برعکس این روش کار می کند به این صورت که بر اساس یک لغت داده شده، می خواهد چند لغت قبل و بعد آنرا تشخیص دهد و با تغییر مداوم اعداد بردارهای لغات، نهایتا به یک وضعیت باثبات می‌رسد که همان بردارهای مورد بحث ماست.

یکی دیگر از الگوریتم هایی که مشابه با روش گوگل، برای نمایش برداری کلمات و ایجاد یک بردار با در نظرگرفتن همجواری کلمات توسط دانشگاه استنفورد پیشنهاد شده است روشGloVe است که کارآیی آن هم تقریباً مشابه با روش Word2Vec ارزیابی شده است اما ما در پژوهش خود از روش اول استفاده‌کردیم به نحوی که بردار کلمات را با تمام لغات موجود در داده بدست آوردیم. ابعاد بردارهای کلماتی که ما با این روش بدست آورده‌ایم 1در50 می‌باشد. زیرا با توجه به کوچک بودن سایز مجموعه داده‌ی آموزشی سایز 50 برای بردارها مناسب است.

4-4- مدل پیشنهادی

4-4-1- نوع مدل

مدلی که در این پژوهش استفاده شده‌است یک شبکه‌ی عصبی کانولوشن با تعداد لایه‌ی کم می‌باشد. در برخی

مسائل، رابطه‌ی بین خروجی و ورودی‌ها بسیار پیچیده هستند و با روش‌های کلاسیک نمی‌توان به سادگی خروجی را بیش‌بینی کرد. یکی از روش‌های استخراج ویژگی از داده‌ی خام، روش‌های یادگیری عمیق است که در این نوع الگوریتم‌ها ما استخراج ویژگی را به دلیل پیچیدگی مسئله به خود الگوریتم می‌سپاریم و هدف بالابردن دقت خروجی پیش‌بینی شده‌است بنابراین ویژگی‌هایی که مدل برای بیش‌بینی استخراج می‌کند لزوما قابل تفسیر نیستند

شبکه های عصبی کانولوشن متشکل از نورون‌هایی با وزن‌ها و بایاس‌های قابل یادگیری هستند. هر نورون تعدادی ورودی دریافت کرده و سپس حاصل ضرب وزن‌ها در ورودی‌ها را محاسبه کرده و در انتها با استفاده از یک یک تابع تبدیل (فعال سازی) غیرخطی نتیجه ای را ارائه دهد. کل شبکه همچنان یک تابع امتیاز مشتق‌پذیر را ارائه می‌کند، که در یک طرف آن داده‌های خام تصویر ورودی و در طرف دیگر آن امتیازات مربوط به هر دسته قرار دارد. این نوع شبکه‌ها هنوز یک تابع هزینه (Loss function) مثل (SVM,Softmax) در لایه آخر تماما مرتبط یا (fully connected) دارند و تمامی نکات مطرحی در مورد شبکه‌های عصبی معمولی در اینجا هم صادق است.

معماری های شبکه های عصبی کانولوشن بصورت صریح فرض میکنند که ورودی های آنها داده‌ی خام است، با این فرض ما می‌توانیم ویژگی‌های مشخصی را درون معماری تعبیه (encode) کنیم .با این عمل تابع پیشرو (forward function) را می‌توان بصورت بهینه‌تر پیاده‌سازی کرد و همینطور با این کار تعداد پارامترهای شبکه نیز بشدت کاهش پیدا می‌کند.

شبکه‌های عصبی یک ورودی دریافت می‌کنند. (در قالب یک بردار که در اینجا ما بردار کلمات را ورودی قرار می‌دهیم) و سپس آنرا از تعدادی لایه مخفی (Hidden layer) عبور می‌دهند. و نهایتا یک خروجی که نتیجه پردازش لایه‌های مخفی است در لایه خروجی شبکه ظاهر می‌شود. هر لایه مخفی از تعدادی نورون تشکیل شده که این نورون ها به تمام نورون‌های لایه قبل از خود متصل می‌شوند. نورون‌های هر لایه بصورت مستقل عمل کرده و هیچ ارتباطی با یکدیگر ندارند .آخرین لایه تماما متصل معمولا نقش نمایش‌دهنده امتیاز هر دسته (class) را ایفا میکند. در اینجا ما کلاس‌ها را به صورت باینری کد کرده‌ایم به این صورت که برای مثال درمدلی که قرار است بعد I وE را پیش‌بینی کند، کد کلاس I به صورت (0و1) و کلاس E به صورت (1و0) می‌باشد. شبکه‌های کانولوشن معماری شبکه را به روش معقولی محدود کردند. بطور خاص، برخلاف یک شبکه عصبی معمولی، لایه های یک شبکه عصبی کانولوشن به اختصار (ConvNet) شامل نورون‌هایی است که در سه بعد عرض، ارتفاع و عمق قرار گرفته اند (مرتب شده اند). کلمه عمق در اینجا اشاره به بُعد سوم

یک توده فعال سازی (activation volume) دارد و به معنای عمق یک شبکه عصبی کامل که به معنای تعداد لایه های موجود در آن است نمی‌باشد. هر نورون در هر لایه بجای اتصال با تمام نورون‌ها در لایه قبل تنها به ناحیه کوچکی از لایه قبل از خود متصل است.

یک شبکه عصبی کانولوشن (ConvNet) نورون های خود را در ۳ بعد مرتب می‌کند (عرض ، ارتفاع و عمق) هر لایه یک شبکه ConvNet ورودی را در قالب یک توده سه بعدی به یک توده سه بعدی خروجی از مقادیر فعال‌سازی نورون‌ها تبدیل می‌کند. در نتیجه یک شبکه ConvNet از چند لایه تشکیل می‌شود و هر لایه شیوه کار ساده‌ای دارد. که در آن یک توده سه بعدی ورودی دریافت کرده و آن را با استفاده از توابعی مشتق پذیر (differentiable function) که ممکن است با پارامتر یا بدون پارامتر باشند به یک توده سه بعدی خروجی تبدیل می‌کند.

آنجایی که مقادیر مربوط به این پارامترهای مراحلی بصورت خودکار تنظیم می‌شود، ما از ان به یادگیری یاد می‌کنیم ، چرا که شبکه عصبی گام بگام با یادگیری این پارامترها قادر به انجام وظیفه شناسایی محول شده به آن می‌شود.

هر لایه شبکه کانولوشن یک توده فعال‌سازی را از طریق یک تابع مشتق‌پذیر به توده فعال‌سازی دیگر تبدیل می‌کند. ما از سه نوع اصلی لایه ها برای ساخت یک معماری شبکه کانولوشن استفاده میکنیم .

این لایه ها عبارتند از : لایه کانولوشن، لایه Pooling و لایه تماما متصل (Fully connected layer) که دقیقا همانند همان که در شبکه‌های عصبی معمولی میبینیم است . ما این لایه‌ها را روی هم قرار می‌دهیم تا یک معماری کامل از شبکه کانولوشن ایجاد کنیم.

4-4-2- ساختار مدل

پژوهشگران به وسیله‌ی بررسی کلماتی که افراد استفاده می‌کنند، درک بهتری از روانشناسی انسانی بدست آورده‌اند. روش رایج آنالیز زبان شامل شمارش کلمات استفاده شده در یک دسته‌بندی از قبل مشخص شده می‌باشد. در این راستا پژوهشگران کلمات را در 64 دسته تقسیم می‌کنند. این دسته بندی واژگان که بسیار استفاده میشود تحقیق زبانی و تعداد کلمات (Linguistic Inquiry and Word Count) یا به اختصار LIWC نامیده می‌شود (Pennebaker JW, 2007). در این روش تعداد کلماتی که افراد در دسته‌های مختلف استفاده می‌کنند شمارش می‌شوند. این روش که به روش واژگان بسته[1] شناخته می‌شود در واقع از یک دانش قبلی در مورد زبان، برای آنالیز متون استفاده می‌کند. این دسته بندی‌ها وابسته به متن هستند. بنابراین دسته‌بندی‌ای که از روی متون علمی بدست آید با دسته‌بندی متون عادی تفاوت خواهد داشت. زیرا کلمات در متون مختلف

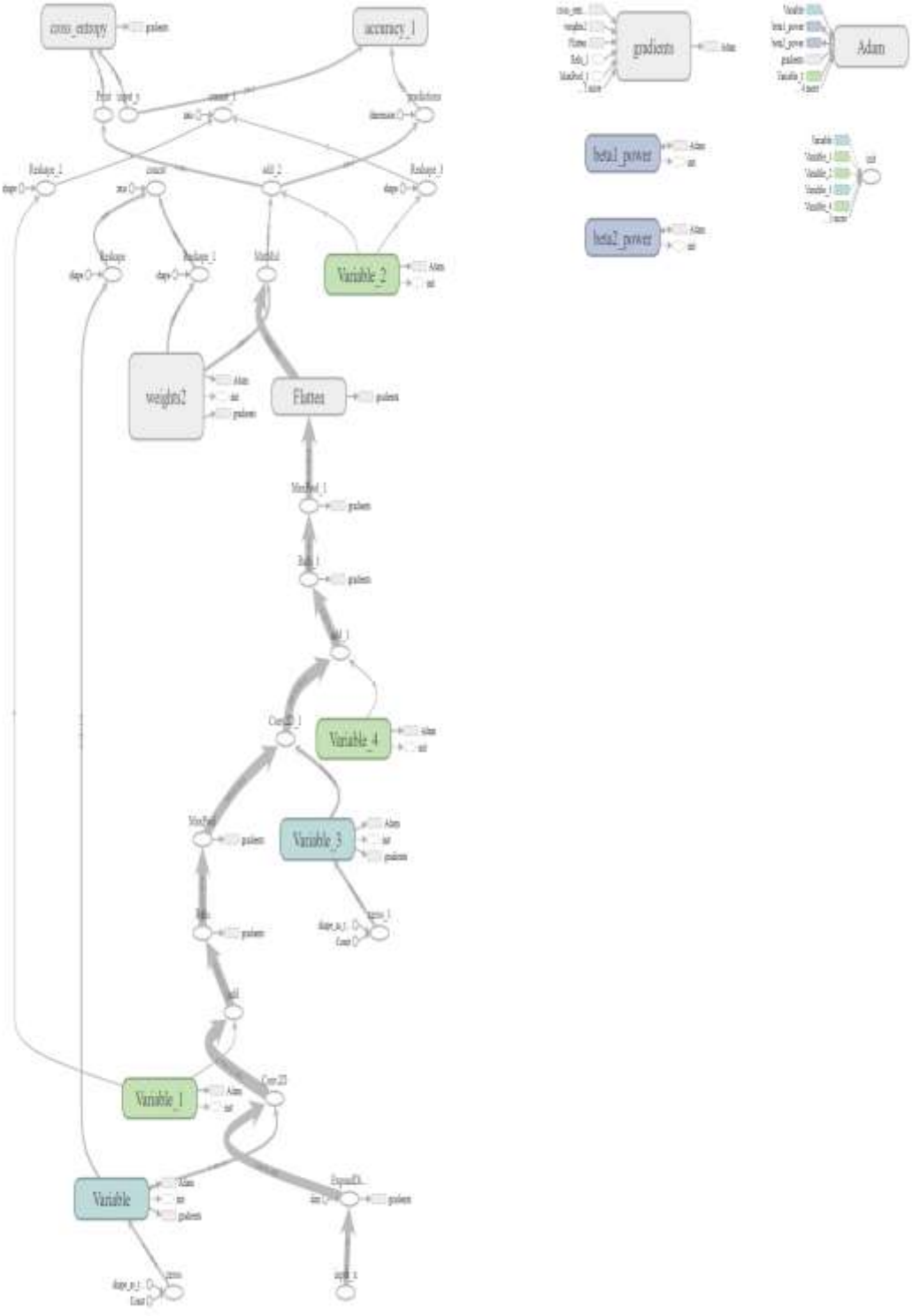
[1] Close-vocabulary

با الگوهای مختلفی استفاده می‌شوند. الگوریتم‌های کلاسیک دیگری نیز از قبیل الگوریتم TF-IDF برای استخراج ویژگی از متن بر اساس تعداد تکرار کلمات وجود دارند. اما این الگوریتم‌ها وابسته به داده‌ای که برای آنها استفاده شده است هستند و ما به دنبال روش‌های عمومی‌تری برای این کار هستیم.

امروزه استفاده‌ی فراگیر از فضای مجازی منجر به این شده‌است که داده‌ی زیادی در اختیار داشته باشیم که بخش اعظم این داده شامل تعاملات روزمره‌ی افراد با یکدیگر است. همچنین استفاده از جملات تک کلمه‌ای نیز بسیار رایج شده‌است بنابراین می‌توانیم رابطه‌ی تک کلمات را با شخصیت افراد پیدا کنیم. اما روابطی بین کلمات وجود دارد که در دسته‌بندی‌ها حضور ندارند و این دسته بندی‌ها محدودیت روی مسئله میگذارند. بنابراین بهتر خواهد بود که بجای محدود کردن مسئله به چند ویژگی خاص، از خود متن ویژگی استخراج کنیم.

استخراج ویژگی از خود متن به روش واژگان باز[1] شناخته می‌شود. پژوهش‌هایی که از این روش استفاده کرده‌اند ویژگی‌های بهتری را نسبت به روش قبلی برای پیش بینی شخصیت یافته اند و این مسئله موجب شد تا ما این روش را برای استخراج ویژگی از زبان برای تعیین شخصیت افراد انتخاب کنیم.

[1] Open-vocabulary

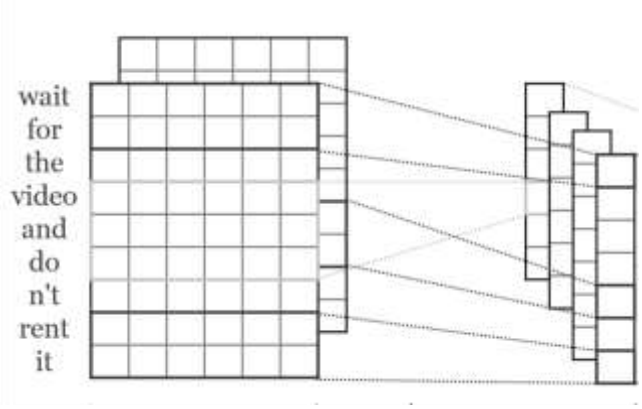


شکل4-2 ساختار شبکه عصبی استفاده شده در پژوهش

در شکل 4-2 ساختار شبکه عصبی استفاده شده در این پژوهش نشان داده شده‌است. همانطور که در شکل مشخص است شبکه ورودی‌ها که شامل 200 پست از افراد مختلف که هرکدام به طور میانگین 12 کلمه در سایز برداری 50 هستند به شبکه وارد می‌شوند. به‌طور میانگین در هر پست 12 کلمه در مجموعه داده وجود دارد. پست‌هایی که بیشتر از این تعداد کلمه دارند، فقط 12 کلمه‌ی ابتدای آنها به شبکه وارد می شوند و بقیه دور ریخته می‌شوند. همچنین پست‌هایی که کمتر از 12 کلمه هستند، جایگاه‌های خالی با بردارهای 50 بعدی که در همه‌ی بعدها عدد یک دارد، پر می‌شوند و در نهایت داده‌ی ورودی با ابعاد 200×12×50 وارد سیستم می‌شود.در مرحله‌ی بعد برای اینکه داده وارد لایه‌ی conv2D شود باید 4 بعدی باشد بنابراین ابعاد داده را به 200×12×50×1 گسترش می‌دهیم. در اینجا سایز پنجره‌ای که وزن‌ها در آن ثابت است و روی داده‌ی 12در50 حرکت می کند 5 در50 است. یعنی ما فرض کرده‌ایم 5 کلمه‌ی متوالی در یک متن ویژگی‌های متن را در بر دارد. سپس این پنجره روی همه‌ی 200 داده اعمال می‌شود. داده‌ها در این مرحله با وزن‌هایی که در ابتدا به صورت رندم از توزیع نرمال با میانگین صفر و انحراف معیار یک تولید شده‌اند ضرب می‌شوند و نتیجه با بایاسی که به در ابتدا با مقادیر صفر تعریف شده‌اند جمع می‌شوند و از تابع فعال‌ساز Relu عبور می‌کنند. سپس داده ها از 1 کانال که بعد اول داده‌ها بود به 4 کانال مختلف وارد می‌شوند تا هرکدام از این کانال‌ها برای مرحله‌ی بعد یک ویژگی s را با تنظیم کردن وزن‌های پنجره بدست بیاورند.

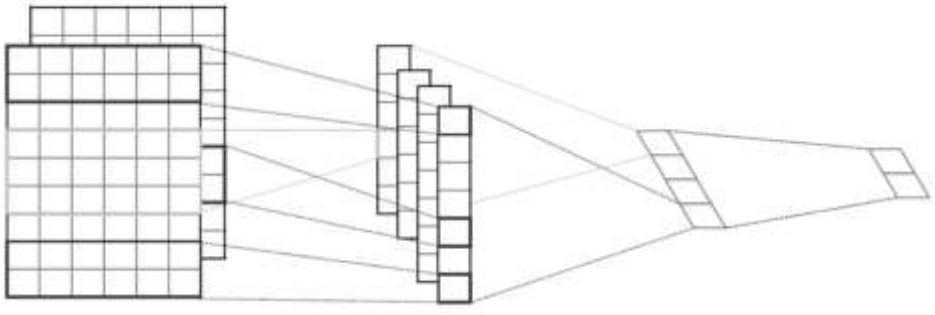
خروجی این مرحله ورودی لایه‌ی بعد که maxpool می‌باشد، است. در این لایه تابع maxpool عملیات خود را با سایز فیلتر 5 تایی انجام می‌دهد و همه‌ی کانال‌ها را به عنوان ورودی به لایه‌ی conv2D بعدی می‌دهد. در لایه‌ی بعد نیز مشابه conv2D قبل عملیات انجام می‌شود با این تفاوت که خروجی 6 کانال دارد و سایز پنجره‌ی تابع maxpool 3 می‌باشد.

در نهایت خروجی لایه‌ی قبل وارد یک لایه‌ی flatten می‌شود. در این لایه تمام کانال‌ها کنار یکدیگر قرار می‌گیرند و ابعاد کم می‌شود. پس از آن وارد لایه‌ی تماما متصل می‌شویم که خروجی آن دو بعدی است و به ما برچسب کلاس پیش‌بینی شده را می‌دهد.



شکل4-3 یک لایه کانولوشن و maxpool

در شکل 4-3 به صورت تقریبی در مقیاس کوچک دو لایه‌ی اول نشان داده‌شده‌است. در لایه‌ی اول ورودی قرار می‌گیرد به این صورت که بردار کلمات در سطرها قرار می‌گیرند بنابراین 12 سطر و 50 ستون داریم و همین ماتریس برای 200 داده به همین صورت تکرار می‌شود. سپس پنجره‌ای به اندازه‌ی 5 سطر و 50 ستون به تعداد کانال‌های خروجی تعریف شده روی ورودی حرکت می‌کند و مقادیر وزن‌های آن در مقادیر بردارهای ورودی ضرب می‌شود و با یک مقدار ثابت بایاس جمع می‌شود و مجموع با استفاده از یک تابع فعال‌ساز به لایه‌ی بعد وارد می‌شود.



شکل4-4 ساختار کلی مدل CNN

شکل 4-4 یک شمای تقریبی از لایه‌های انتهایی مدل را نشان می‌دهد. در لایه‌ی اول که کانولوشن را نشان می‌دهد، جدول‌های مختلف نشان‌دهنده‌ی کانال‌های متفاوت است که هرکدام حاصل یک پنجره وزن متفاوت از مرحله‌ی قبل هستند. سپس پنجره وزن این لایه به تعداد تعریف شده روی داده‌ها حرکت می‌کنند و به لایه‌ی maxpool می‌روند. پس از آن خروجی‌های این لایه در flatten کنار یکدیگر قرار می‌گیرند تا به لایه‌ی fully connected وارد شوند و خروجی را تولید کنند.

برای منظم سازی از dropout با احتمال 0.25 استفاده شده است همچنین از adam optimizer با نرخ یادگیری 0.001 نیز برای بهبود نتایج حین آموزش استفاده شده است. به‌روز رسانی وزن‌ها و بایاس‌ها در همه‌ی لایه ها توسط تابع بهینه‌ساز adam که از Stochastic Gradient Descent استفاده می‌کند، انجام می‌شود. در واقع شبکه با ورودی X و خروجی Y به صورت زیر از فرمول‌های 4-1 تا 4-8 عمل می‌کند. که در نهایت شبکه با محاسبه‌ی میزان تفاوت خروجی شبکه و خروجی واقعی T دقت و هزینه را محاسبه و پارامترها را با الگوریتم adam بروزرسانی می‌کند.

$$\text{Conv2D}(W_0X + B_0)=Y_0 \quad (4\text{-}1)$$

$$\text{ReLu}(Y_0)= Z_0 \quad (4\text{-}2)$$

$$\text{Maxpool}(Z_0)= Y_1 \quad (4\text{-}3)$$

$$\text{Conv2D}(W_1Y_1 + B_1)= Y_2 \quad (4\text{-}4)$$

$$\text{ReLu}(Y_2)= Z_1 \quad (4\text{-}5)$$

$$\text{Maxpool}(Z_1)= Y_2 \quad (4\text{-}6)$$

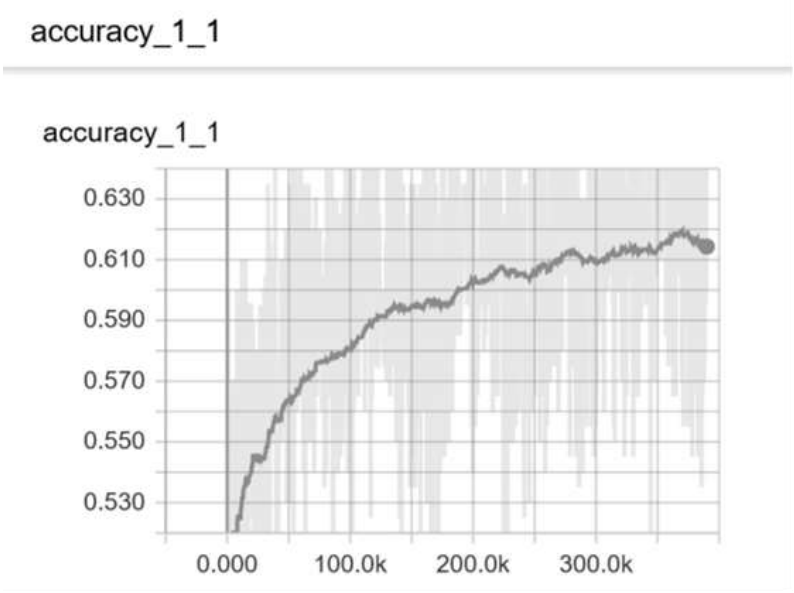
$$\text{Flatten}(Y_2)= Y_3 \quad (4\text{-}7)$$

$$\text{FullyConnected}(W_2Y_3 + B_2) =\text{Y} \quad (4\text{-}8)$$

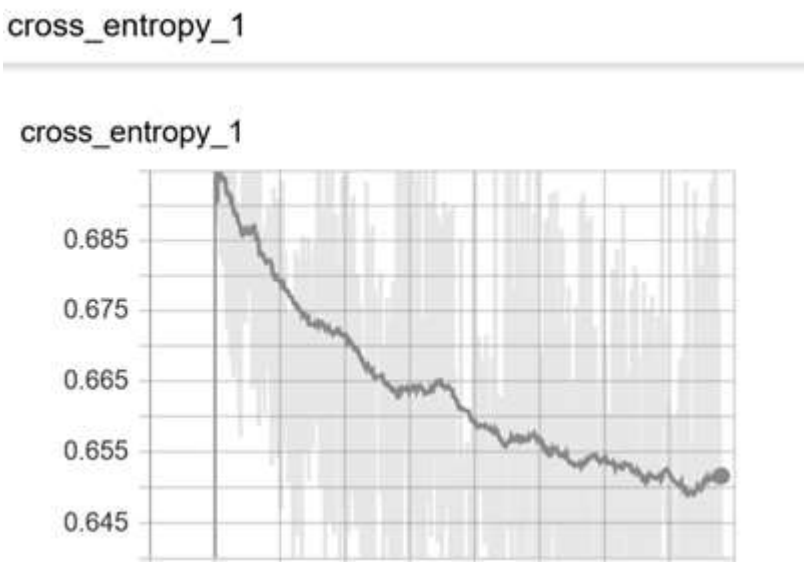
4-5- نتایج

ما دو فاکتور شخصیتی را برای در نظر گرفتیم و دو مدل را برای پیش‌بینی آنها آموزش دادیم. یک مدل برای پیش‌بینی بعد درونگرا- برونگرایی و یک مدل برای پیش‌بینی بعد شمی- حسی در نظر گرفتیم. دو مدل به لحاظ ساختار و تعارف اولیه کاملا شبیه به یکدیگر هستند. و فقط داده‌ی آنها با هم متفاوت است.

برای مدل اول همه‌ی تیپ‌های شخصیتی که درونگرا بودند (شامل هشت تیپ) را یک کلاس در نظر گرفتیم و همه‌ی تیپ‌های شخصیتی که برونگرا هستند کلاس دیگر را تشکیل می‌دهند. همین کار را برای مدل دوم انجام داده‌ایم. هر دو شبکه را در شرایط یکسان در 500 تکرار آموزش دادیم و دقت و هزینه (loss) مدل را که با تابع cross entropy محاسبه می‌شود در فاز تست با 50000 داده که سیستم در مرحله‌ی آموزش ندیده است، اندازه گرفتیم.در این 50000 داده توزیع داده‌ها در کلاس‌ها یکسان نیستند زیرا لزومی ندارد که در مرحله‌ی تست توزیع داده‌ها یکسان باشند و ما یک نمونه از جامعه‌ی واقعی را برای آزمایش به مدل داده‌ایم.

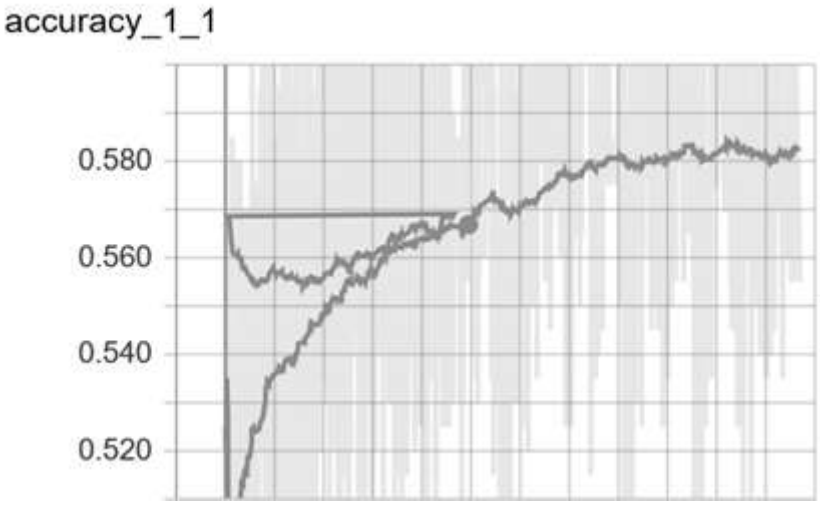


شکل 4-5 نمودار دقت مدلI-E

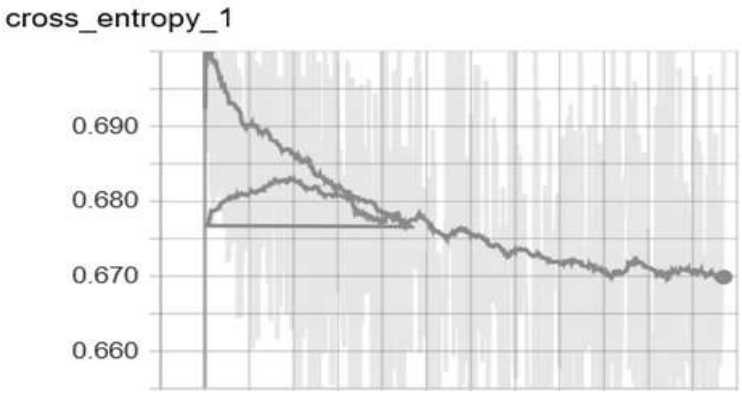


شکل 4-6 نمودار هزینه‌ی مدلI-E

همانطور که در شکل‌های 4-5 و 4-4 مشخص است دقت این مدل در انتهای آموزش به 62٪ و هزینه‌ی مدل در انتهای آموزش به 0.64 رسیده است.



شکل 4-7 نمودار دقت مدل N-S



شکل 4-8 نمودار هزینه‌ی مدل N-S

همانطور که در شکل‌های 4-6 و 4-7 مشخص است دقت این مدل در انتهای آموزش به 58٪ و هزینه‌ی مدل در انتهای آموزش به 0.67 رسیده است.

فصل پنجم:

نتیجه گیری و پیشنهادات

5-1- مقدمه

در این فصل که فصل پایانی این پایان‌نامه می‌باشد به بررسی و تحلیل نتایج بدست آمده می‌پردازیم. در فصل گذشته نتایج حاصل از شبکه‌ی عصبی طراحی شده برای 2 بعد از شخصیت افراد را بر مبنای تست MBTI از روی متن آنها را دیدیم. تلاش بر این است که در این فصل به آنالیز و مقایسه‌ی مسئله با مسائل مشابه در این حوزه بپردازیم و جنبه‌های مثبت و منفی و نقات قوت و ضعف و چالشی مسئله را بیان کنیم. همچنین بیان خواهیم کرد که در این پژوهش ما توانستیم چه قسمت‌هایی از مسئله را بپوشانیم.

5-2- تشریح نتایج و یافته‌های تحقیق

همانطور که در فصل پیش اشاره کردیم توزیع داده‌ها در کلاس‌های شخصیتی یکسان نیست و توزیع آن‌ها در شکل 1-4 نشان داده‌شده‌است و ما برای آموزش توزیع آن‌ها را در کلاس‌های باینری یکسان نمودیم به این صورت که ابتدا داده‌ها را بر اساس این که برچسب آن‌ها I یا E است جدا کردیم و یک مجموعه داده با دو برچسب I و E به دست آوردیم و سپس توزیع داده را در دو کلاس یکسان کردیم و داده را برای آموزش آماده کردیم همین کار برای تقسیم داده‌ها با برچسب N و S انجام دادیم. بنابراین ما دو مجموعه داده متفاوت که براساس کلاس‌هایشان به دست‌آمده‌اند داریم. این کار را برای این که در فرایند آموزش شانس هر کلاس مساوی باشد انجام دادیم اما در فاز تست توزیع دادها را یکسان نمی‌کنیم زیرا برای تست فرض می‌کنیم که از یک جامعه‌ی واقعی نمونه گرفته‌ایم.

با یک بررسی در مجموعه داده‌ها متوجه شدیم که توزیع داده‌ها به صورت چشمگیری با یکدیگر متفاوت هستند برای مثال از افراد درونگرا داده‌ی بیشتری در اختیار داشتیم و از برخی از تیپ‌های خاص شخصیتی نیز داده‌ی بیشتری در دسترس بود. این نشان می‌دهد که افراد درونگرا احتمالا تمایل بیشتری به استفاده از فضای مجازی دارند و یا احتمالا بیشتر نظرات خود را در فضای مجازی به اشتراک می‌گذارند. همچنین مشاهده کردیم که طول متنی که از افراد برونگرا در اختیار داریم نسبت به افراد درونگرا کمتر است و این افراد بیشتر تمایل به استفاده از جملات کوتاه دارند.

با یک بررسی دیگر در مجموعه داده متوجه شدیم که افراد درونگرا تمایل به استفاده از کلمات و جملات پیچیده‌تری دارند و نظرات، احساسات و افکار خود را به روش پیچیده‌تری بیان می‌کنند و این با نتایجی که پژوهشگران در این زمینه از قبل بدست‌آورده‌بودند هم‌خوان بود.

ما می‌خواستیم که وجود داده‌ی زیاد در افراد درونگرا بر تست تاثیری نگذارد اما در واقعیت این‌طور است که تشخیص افراد درونگرا از روی متن ساده‌تر است زیرا هم داده‌ی بیشتری از آن‌ها در اختیار است و هم اکثرا

نوع نوشتاری خاصی دارند اما در شرایط برابر که احتمال هر دو کلاس مساوی است سیستم باید بتواند با احتمالی مناسب داده را به یک کلاس اختصاص دهد.

پژوهشگران تلاش‌های زیادی در این زمینه انجام داده‌اند که بتوانند رابطه‌ای بین شخصیت و متنی که افراد منتشر می‌کنند به دست بیاورند اما در این پژوهش هدف صرفا تشخیص تیپ شخصیتی از روی متن است و فقط خروجی برای ما اهمیت دارد. در پژوهش‌های گذشته برای این کار که بر روی تست Big-Five انجام شده‌است درصد دقت خوبی بدست آمده‌است. اما این تست برای هر کلاس طیفی از صفر تا 100 در نظر می‌گیرد و مسئله به رگرسیون تبدیل می‌شود که بسیار ساده‌تر است اما در تست MBTI که اطلاعات ریزتر و جزئیات بیشتری را در مورد شخصیت می‌دهد مسئله کلاس‌بندی است و این مسئله با مسئله‌های کلاس بندی متون دیگر مقایسه می‌شود. در زمینه‌ی کلاس بندی متون نیز تلاش‌هایی شده‌است که منجر به نتایج خوب شدند اما مسئله‌ی شخصیت بسیار پیچیده‌تر می‌باشد به این صورت که در بعضی افراد هنوز بعضی از ابعاد شخصیتی ثابت نشده‌اند و آن‌ها در جایی بین دو کلاس و کمی بیشتر متمایل به یک کلاس خاص قرار دارند. این مسئله به اینکه چقدر شخصیت انسان در او تثبیت شده‌باشد برمی‌گردد که از لحاظ روانشناسی به فاکتورهایی مانند سن و غیره وابسته است. ولی متاسفانه ما در مجموعه داده خود اطلاعاتی از سن و دیگر ویژگی‌های این افراد در دسترس نداشتیم و شاید برای یک فرد متخصص روانشناسی هم تشخیص تیپ شخصیتی فردی که هنوز شخصیت در او تثبیت نشده‌است از روی یک متن کوتاه دشوار باشد اما اگر شخصیت تثبیت شده‌باشد با دقت بالایی می‌توان این کار را انجام داد. بنابراین پیچیدگی مسئله بسیار بالا است و ما شبکه‌ی عصبی عمیق را برای این کار انتخاب کردیم تا نتایج نسبتا خوبی بدست‌آوریم.

همه‌ی ابعاد شخصیتی با یکدیگر یکسان نیستند و در برخی تشخیص آن‌ها بسیار دشوارتر است همانطور که نتایج نشان می‌دهند، تشخیص این که یک فرد درونگرا یا برونگرا است از این که فردی شمی یا حسی است راحت‌تر است با اینکه دو شبکه و ساختار و همه چیز با یکدیگر برابر بودند و در شرایطی برابر دو شبکه در حال آموزش بودند. البته این می‌توانست نتیجه‌ی کمبود داده نیز باشد که ما با روش‌های منظم‌سازی سعی کردیم با این کمبودا مقابله کنیم و مدل را عمومی‌تر بسازیم.

این یافته می‌تواند برای علاقه‌مندان به این حوزه جالب و قابل تامل باشد که در بین بعدهای شخصیتی براساس مدل MBTI تشخیص بعد اول از دوم از روی متن ساده‌تر است.

5-3- خلاصه، بحث و تبیین نتایج

مسئله‌ی تعیین تیپ شخصیتی افراد در حوزه‌ی مسائل آنالیز متن قرار می‌گیرد. طبقه بندی متن، یعنی انتساب اسناد متنی بر اساس محتوی به یک یا چند طبقه از قبل تعیین شده، یکی از مهمترین مسائل در متن کاوی

است. متن کاوی به دنبال استخراج اطلاعات مفید از داده های متنی غیر ساخت‌یافته از طریق تشخیص و نمایش الگوها است یا به عبارت دیگر متن کاوی روشی برای استخراج دانش از متون است. متن کاوی کشف اطلاعات جدید و از پیش ناشناخته، به وسیله استخراج خودکار اطلاعات از منابع مختلف نوشتاری است.
داده کاوی در متن در زمان‌های مختلف بر اساس کاربرد و روش شناسی مورد استفاده، به صورت پردازش متن آماری، کشف دانش در متن، آنالیز هوشمند متن یا پردازش زبان طبیعی تعیین شده است. به عنوان مثال هایی از کارهایی که متن کاوی انجام می دهد می توان به دسته بندی یا classifying اسناد به مجموعه‌ای از تاپیک های مشخص (یادگیری با نظارت) که مسئله‌ی ما نیز در همین زمینه قرار دارد اشاره کرد.
متن کاوی به عنوان تجزیه و تحلیل هوشمند متن، داده کاوی متن یا کشف دانش در متن نیز شناخته می‌شود. متن کاوی بر روی داده‌های متنی غیرساخت‌یافته و نیمه ساخت‌یافته تعریف می‌گردد داده‌های متنی غیرساخت یافته مانند صفحات وب، یادداشت، صورتحساب و غیره هستند که داده‌ی مورد استفاده در این پژوهش نیز جور داده‌های غیر ساخت یافته قرار می‌گیرد.
کاوش داده های غیر ساخت‌یافته با پردازش زبان طبیعی(NLP) ، مدلسازی آماری و روش های یادگیری ماشین ممکن است سخت و چالش برانگیز باشد چون متن‌های زبان طبیعی اغلب متناقض هستند. این متن ها اغلب شامل ابهاماتی هستند که از سینتکس‌ها و معناشناسی‌های (سمنتک) متناقض مانند اصطلاحات عامیانه یا زبان‌های مربوط به یک گروه سنی خاص یاصحبت‌های کنایه‌دار و طعنه‌آمیز نشات می‌گیرد.
به طور کلی روش‌هایی که در متن کاوی استفاده می‌شوند عبارتند از:
استخراج اطلاعات، طبقه بندی، خوشه بندی، خلاصه سازی، ردیابی موضوع، ارتباط دهنده مفاهیم، نمایش اطلاعات، پرسش و پاسخ، کاوش مبتنی بر متن، تجزیه و تحلیل گرایش‌ها.

5-3-1- استخراج اطلاعات

در استخراج اطلاعات، عبارات کلیدی و ارتباط آنها در متن تشخیص داده می‌شود این عمل بوسیله پردازش تطبیق دهنده الگو انجام می‌پذیرد و عبارات و اصطلاحات استخراج شده باید بصورت استاندارد باشد.

5-3-1-1- استخراج ویژگی

اولین مرحله متن کاوی استخراج ویژگی یا feature extraction در مجموعه اسناد است به طوری که شخص بتواند محاسبات انجام داده و از روش های آماری استفاده کند.
در متن کاوی از دو کلمه corpus و lexicon استفاده می شود که corpus به معنی مجموعه ای از اسناد است و بسیاری از روش‌های استخراج ویژگی وابسته به corpus هستند و lexicon یا واژه نامه مجموعه‌ای از همه کلمات منحصر به فرد در corpus است.

5-3-1-2- تولید و انتخاب ویژگی

اسناد متنی توسط لغات (ویژگی‌هایی) که دارند و ارتباط میان آنها نمایش داده می شوند دو رویکرد عمده نمایش اسناد bag of words و Vector Space هستند. که ما از دویکرد دوم با الگوریتم wor2vec استفاده کرده‌ایم.

از یک دسته بندی کننده یا classifier برای تولید خودکار لیبل ها (ویژگی ها) از ویژگی هایی که به آن داده شده استفاده می کند. در روش‌های یادگیری ژرف وظیفه‌ی استخراج ویژگی را به شبکه می‌سپاریم و لزوما ویژگی‌های استخراج شده را تفسیر نمی‌کنیم.

5-3-2- طبقه بندی

هدف از طبقه‌بندی، ایجاد امکان استفاده از مدلی بر ای پیش‌بینی کلاسی از اشیا است که با عنوان ناشناخته برچسب خورده است.

طبقه بندی یک فرایند ۲ مرحله‌ای است:

الف- ساخت مدل

ب- استفاده از مدل

در توسعه دسته بندی برای اسناد متنی چالش هایی وجود دارد مثلا یکی از این چالش ها برخورد با مترادف ها و کلمات چند معنی است. چالش دیگر ایجاد دسته‌بندی‌هایی است که بتواند مجوعه‌های بزرگ اسناد را دسته‌بندی کند. یا چالش دیگر دسته بندی منابع اسناد در حال استریم است. مانند اخبار که بصورت مداوم پخش می‌شوند.

هدف از طبقه بندی متون نسبت دادن کلاس‌های از پیش تعریف شده به اسناد متنی است. در طبقه‌بندی یک مجموعه آموزشی از اسناد، با کلاس‌های معین وجود دارد .با استفاده از این مجموعه، مدل طبقه‌بندی معین شده و کلاس سند جدید مشخص می‌گردد. برای اندازه گیری کارایی مدل طبقه‌بندی، یک مجموعه تست، مستقل از مجموعه آموزشی در نظر گرفته می‌شود. برچسب‌های تخمین زده شده با برچسب واقعی اسناد مقایسه می‌شود. نسبت اسنادی که به درستی طبقه‌بندی شده‌اند به تعداد کل اسناد، دقت نامیده می‌شود.

5-3-3- شبکه های عصبی

در مسائل مربوط به طبقه‌بندی، شبکه عصبی با داشتن ورودی‌ها و خروجی‌های مشخص باید تشخیص دهد که هر ورودی با کدام طبقه از خروجی‌های تعریف شده بیشترین تطابق را دارد . هدف از آموزش شبکه به حداقل رساندن خطای تولید شده می‌باشد که براساس تنظیم وزن‌های شبکه انجام می‌شود. معمولا از الگوریتم

آموزش پس‌انتشار استفاده می‌شود. پس از محاسبه مقدار خطا در لایه خروجی مقادیر وزنها در لایه پنهان در جهت کاهش خطا تنظیم می‌شوند.

اقتباس واژه، اساسی‌ترین شکل متن‌کاوی است. مانند تمام تکنیک‌های دیگر متن‌کاوی اطلاعات را از داده ساخت‌نیافته به یک فرمت ساخته‌یافته نگاشت می‌دهد. ساده‌ترین ساختمان‌داده در متن‌کاوی، بردار ویژگی یا لیست وزن‌دار کلمات است. مهم‌ترین کلمات در یک متن به همراه اندازه اهمیت نسبی آن ها فهرست می شود. متن به فهرستی از واژگان و وزن‌ها کاهش می‌یابد. کل معنا شناختی یک متن ممکن است وجود نداشته باشد، ولی مفاهیم کلیدی شناسایی می‌شوند.

داده‌کاوی، بازیابی اطلاعات، یادگیری ماشین، پردازش زبان طبیعی و استخراج اطلاعات از زمینه های مرتبط با متن‌کاوی هستند. این تکنیک‌ها به همراه هم برای کشف خودکار الگوها در اطلاعات استخراج شده و متادیتای بدست آمده از مستندات بکار می‌روند.

می‌توان گفت که متن‌کاوی از تکنیک‌های بازیابی اطلاعات، استخراج اطلاعات همچنین پردازش‌کردن زبان طبیعی استفاده کرده و آن‌ها را به الگوریتم‌ها و متدهای KDD داده‌کاوی، یادگیری ماشین و آماری مرتبط می‌کند.

ما در این پژوهش توانستیم مدلی مبتنی بر یادگیری ژرف برای تشخیص شخصیت بر مبنای MBTI ارائه دهیم که تاکنون انجام نشده بود. کارهای گذشته بر روی مدل‌های شخصیتی دیگر انجام گرفتند که برخی از ویژگی-های از پیش تعیین شده‌ی زبانی برای پیش‌بینی استفاده کرده‌اند اما ما بدون استفاده از دانش پیشین توانستیم برای این مسئله مدل نسبتا خوبی ارائه دهیم و نتایج را از تنها پژوهشی که دقیقا مشابه با این پژوهش در دانشگاه استنفورد انجام شده بود بهبود دهیم. پژوهشگران تا کنون برای این مدل دقتی بالاتر از 50 درصد نیافته اند و ما نشان دادیم که یک مدل شبکه عصبی ژرف می‌تواند دقت را بالا ببرد.

مسئله‌های دیگری که در حوزه‌ی طبقه‌بندی متن تعریف می‌شوند از مسئله تشخیص شخصیت آسان‌تر می‌باشد زیرا این مسئله به خودی خود اگر توسط هوش طبیعی انسان حل شود نیاز به یک دانش روانشناسی و تخصص دارد.

یادگیری ژرف در پردازش تصویر بسیار کاربرد دارد و مفید می‌باشد. در این پژوهش ما کاربرد این نوع شبکه‌های عصبی را که ویژگی‌های پیچیده از داده‌ی خام استخراج می‌کنند در مسئله‌های پیچیده از جمله تشخیص شخصیت نیز نشان دادیم.

اخیرا یادگیری ژرف در طبقه بندی متون نیز استفاده می‌شود که برای مثال به آنالیز نظرات کاربران در باره‌ی یک محصول، مشخص کردن بار مثبت یا منفی یک جمله و غیره می‌توان اشاره کرد. همه‌ی این مسائل مسائلی هستند که هوش طبیعی انسان بدون داشتن اطلاعات تخصصی می‌تواند آنها را حل کند. اما اینکه آیا هوش

مصنوعی می‌تواند در زمینه‌های تخصصی نیز صاحب نظر باشد و نه تنها مانند یک هوش طبیعی بلکه یک هوش طبیعی متخصص نظر بدهد و پیش‌بینی کند.

5-4- پیشنهادات

از آنجایی که در این زمینه داده‌ی بسیار کمی در اختیار می‌باشد جمع‌آوری و اشتراک داده می‌تواند بسیار به محققان در این زمینه کمک کند. همچنین جالب‌تر خواهد شد اگر داده‌ای به زبان فارسی برای این کار داشته باشیم زیرا هیچ داده‌ای موجود نمی‌باشد در این صورت می‌توانیم تخمین بهتری از شخصیت افراد با توجه به متن آن‌ها بدهیم.

همچنین به تازگی پژوهشگران روش character level را برای حل مسائل طبقه‌بندی و دسته‌بندی متون و به طور کلی آنالیز متن پیشنهاد کرده‌اند. که به این صورت است که به جای این که کلمات را بازنمایی کنیم حروف بازنمایی می‌شوند. بررسی این موضوع نیز می‌تواند نتایج جالبی برای محققان داشته باشد.

ما به دلیل کمبود داده و کوچک شدن مجموعه داده برای آموزش نمی‌توانستیم در ابعاد دیگر شخصیتی شبکه را آموزش دهیم. اگر راه‌حلی برای مسئله یافت شود می‌توانیم برای آن ابعاد شخصیتی نیز مدلی ارائه کنیم.

اگر ما فاکتورهای دیگری از افرادی که اطلاعات متن و تیپ شخصیتی آن‌ها را در اختیار داریم می‌داشتیم می‌توانست به تست کمک کند و شاید دید بهتری از مسئله به ما بدهد زیرا در مسئله‌ی شخصیت فاکتورهای زیادی تاثیرگذار هستند که نمی‌توان آن‌ها را نادید گرفت. کاری که ما انجام داده‌ایم بدون در نظر گرفتن هیچ فاکتوری این مسئله را حل می‌کند اما برای دقت بیشتر اطلاعات بیشتری نیاز است.

ما در پژوهش خود از هر فرد 50 پست متنی در اختیار داشتیم و آن پست‌ها را جدا کردیم و فقط با یک پست شخصیت را پیش‌بینی کردیم نتیجه‌ی این که اگر متن بزرگتری و کلمات بیشتری از یک فرد در اختیار داشته باشیم آیا می‌تواند درصد دقت مسئله را بالا ببرد یا خیر نیز از لحاظ روانشناسی دارای ارزش می‌باشد.

در پژوهش‌های پیشین پژوهشگران از شبکه‌های بازگشتی استفاده کرده‌اند ما در پژوهش خود از نوع دیگری از شبکه‌های عصبی به نام CNN استفاده کردیم که استفاده از این شبکه توسط یکی از پژوهشگران صاحب نام در این حوزه توصیه شده بود. مشاهده و مقایسه‌ی این نتایج بدست آمده از این دو شبکه نیز می‌تواند حائز اهمیت باشد.

5-5- محدودیت‌های تحقیق:

بزرگترین و مهم‌ترین محدودیت در پروژه‌هایی از این قبیل که در زمینه‌ی روانشناسی و هوش مصنوعی قرار دارند کمبود داده است. زیرا در پروژه‌ای که ما تعریف کردیم ما از خود متن برای کلاس بندی استفاده می‌کنیم و اغلب داده‌های موجود از ویژگی‌های متنی استفاده می‌کنند و داده‌ای که به صورت متن و تیپ شخصیتی

باشد بسیار نایاب است و تنها داده‌ای که برای انجام این پژوهش در دسترس داشتیم همین داده‌ی مورد استفاده بود.

یک مجموعه داده‌ی بزرگ وجود دارد که اغلب پژوهشگران از آن استفاده می‌کنند و این مجموعه داده مربوط به پروژه‌ی myPersonality که یک اپلیکیشن در فیسبوک است می‌باشد. این مجموعه داده برای عموم در دسترس بود اما متاسفانه این شرکت از سال 2018 تصمیم به عدم اشتراک داده‌ی خود با عموم گرفته‌است.

این کمبود داده پژوهشگران را محدود می‌کند در نتیجه از آن جایی که بهترین داده‌ی موجود در اختیار پژوهشگران قرار نمی‌گیرد پژوهش‌های کمتری انجام خواهد شد. تا پیش از این اغلب پژوهش‌هایی که در این زمینه انجام شده‌بودند بر روی این داده کار کرده‌بودند بنابراین ما علاوه بر مشکل کمبود داده با مشکل کمبود تعداد پژوهش نیز مواجه بودیم.

از آنجایی که این مسئله در حوزه‌ی پردازش زبان‌های طبیعی می‌باشد و این حوزه بسیار زمینه‌ی نوینی در هوش مصنوعی می‌باشد ما بانک اطلاعاتی کوچکی در اختیار داشتیم.

همچنین اجرای این برنامه‌ها و کدها نیازمند سیستم پردازنده‌ی قوی می‌باشد که به صورت موازی پردازش شود زیرا پردازش این اپلیکیشن‌ها بر روی سیستم‌های عادی بسیار زمان بر و هزینه‌بر است و امکان آزمون و خطا را از پژوهشگران می‌گیرد.

منابع و مآخذ

منابع انگلیسی

[1] S. Fortunato, “Community detection in graphs,” *Phys. Rep.*, vol. 486, no. 3–5, pp. 75–174, 2010.

[2] L. Tang, X. Wang, and H. Liu, “Community detection via heterogeneous interaction analysis,” *Data Min. Knowl. Discov.*, vol. 25, no. 1, pp. 1–33, 2012.

[3] A. Rowe, S. Rowe, A. Silverman, and M. L. Borum, “P024 Crohn’S Disease Messaging on Twitter: Who’S Talking?,” *Gastroenterology*, vol. 154, no. 1, pp. S13–S14, 2018.

[4] B. Agarwal, “Personality Detection from Text : A Review,” *Int. J. Comput. Syst.*, vol. 1, no. 1, pp. 1–4, 2014.

[5] Y. Amichai-Hamburger, “Internet and personality,” *Comput. Human Behav.*, vol. 18, no. 1, pp. 1–10, 2002.

[6] R. Ackoff, “Ackoff’s Best,” pp. 170–172, 1999.

[7] R. Samizade, E. Mahmoudi, and S. Abad, “The Application of Machine Learning Algorithms for Text Mining based on Sentiment Analysis Approach,” *J. J. Inf. Technol. Manag.*, vol. 10, no. 102, pp. 309–330, 2018.

[8] E. Younesi, L. Toldo, B. Müller, C. M. Friedrich, N. Novac, A. Scheer, M. Hofmann-Apitius, and J. Fluck, “Mining biomarker information in biomedical literature,” *BMC Med. Inform. Decis. Mak.*, vol. 12, no. 1, 2012.

[9] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, “Predicting

Personality from Twitter.pdf," 2011.

[10] C. C. Doi, "Our Twitter Profiles,Our Selves: Predicting Personality with Twitter," pp. 180–185, 2011.

[11] R. Wald, T. Khoshgoftaar, and C. Sumner, "Machine prediction of personality from Facebook profiles," *Proc. 2012 IEEE 13th Int. Conf. Inf. Reuse Integr. IRI 2012*, pp. 109–115, 2012.

[12] S. Poria, A. Gelbukh, and B. Agarwal, "Advances in Soft Computing and Its Applications," vol. 8266, no. November, 2013.

[13] B. Verhoeven, W. Daelemans, and T. De Smedt, "Ensemble Methods for Personality Recognition," *Proc. Work. Comput. Personal. Recognit. Shar. Task*, pp. 1–4, 2013.

[14] A. Ma and G. Liu, "Neural Networks in Predicting Myers Brigg Personality Type From Writing Style," pp. 1–9, 2017.

[15] A. Essazadegan and D. Ph, " Relationship between the metaphors and Eysenck's introversion-extraversion dimensions," pp. 54–63,.

Abstract

The exponential growth of the use of social networks in cyberspace has led individuals to share a lot of information, including image, voice and text. Analyzing social Networks data provides details information about individual personality. The complexity and large volume of extracted data is that such that it requires to apply machine learning algorithms. In this paper the author has analyzed the behavior patterns using writing. we first need to know the standard personality models. one of the most reliable models is MBTI model. the goal of this thesis is to find a supervised Learning model that can determine personality factors by people writing in social networks. due to the fact that experience has shown for complex problems with many parameters, deep learning methods can be more effective, we used deep learning model and two personality factors are considered an introversion - extroversion and intuition - sensing. The obtained results show a good accuracy which based on we were able to distinguish an introversion - extroversion personality factor with precision of 62 % accuracy and intuition – sensing factor with precision of 58 %

Keywords: Social networks, MBTI, Deep Learning, supervised learning, Machine learnig.

ISLAMIC AZAD UNIVERSITY

SCIENCE AND RESEARCH BRACH

MASTER OF SCIENCE THESIS

COMPUTER ENGINEERING

– SOFTWARE

Subject:

Providing a Model for Detecting Personality of People in Social Networks.

Thesis Advisors:

Mehdi Hosseinzadeh

Peyman Sheikholharam mashhadi

By:

Omid Asghari

autemn 2018