پروژه‌ی تشخیص احساسات در متن

نیکتا گوهری صدر

احساسات نقش مهمی را در شکبه های اجتماعی و میکروبلاگ ها ایفا میکند. با توجه به اینکه بیشتر این داده ها متنی هستند، پیشبینی احساسات از روی متون از اهمیت بالایی برخوردار است. تشخیص این مسئله به وسیله انسان و به روش دستی وقت گیر و هزینه بر است. به همین دلیل تحقیقات زیادی به جهت استخراج این احساسات توسط ماشین از متن ها انجام شده است. در این تحقیق سه نمونه از جدید ترین روش های برای حل این مسئله بررسی و پیاده سازی شده است. هر سه این آزمایشات شامل پیش پردازش، استخراج ویژگی و بردار سازی و پیشبینی احساسات می‌باشد. در آخر نتیجه این سه روش با هم مقایسه و ارزیابی می‌شود.

1. مجموعه داده

طبق آزمایشات انجام شده، دو مدل پیاده سازی شده نتوانستند به طور همزمان هر سطر داده را به حداکثر دو کلاس نسبت دهند. بنابرین داده های دو کلاسه با تکرار سطر تبدیل به یک کلاس شده‌اند که این روش دقت مدل ها را بالا برده است.

2. پیاده‌سازی‌ها

2.1. پیش پردازش

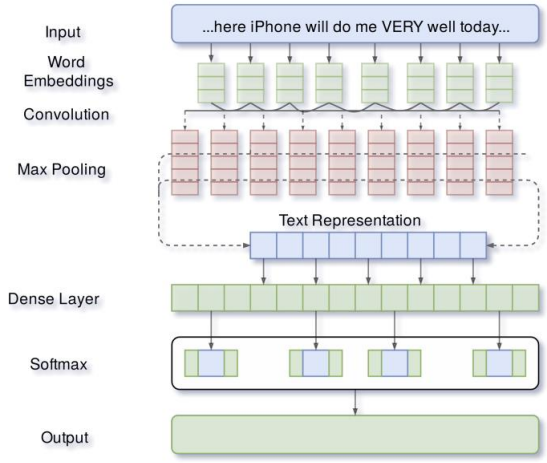
در مرحله‌ی اول برای کم کردن حجم پردازش و بالا بردن دقت، داده های با ارزش نگهداری شده و بقیه حذف شده اند. به همین جهت ابتدا مقدار های خالی و داده های تکراری حذفه شده سپس از متن هر سطر کاراکتر هشتگ، html، حروف تکرار شده و نقطه گذاری ها (به جز علامت سوال و علامت تعجب که برای تشخیص احساسات میتوانند مفید باشند) حذفه می‌شوند. همچنین stop word ها مانند "و"، "در" و غیره حذف شده و کلمات با ریشه اصلیشان جایگزین شده اند.

2.2. آزمایش اول: انتخاب ویژگی با معیار خی دو + نایو بیز چند جمله‌ای

برخی از تحقیقات انجام شده در این زمینه از الگوریتم نایو بیز برای طبقه بندی احساسات متون استفاده کرده اند. [2] به همین دلیل در این تحقیق ابتدا با معیار خی دو 10 هزار ویژگی که همان کلمات داخل متن هستند انتخاب شده و با الگوریتم نایو بیز چند جمله ای این داده ها طبقه بندی شده اند. این آزمایش در دو مرحله انجام شده است. مرحله اول روی داده ها با لیبل هایی حداکثر متعلق به 2 کلاس و در مرحله ی دوم روی مجموعه داده ای که هر سطر متعل به حداکثر یک کلاس است.(تکرار سطر ها) همانطور که انتظار میرود مدل ها با مجموعه داده دوم دقت بیشتری دارند.

2.3. آزمایش دوم: CNN

در سال های اخیر بیشتر از معماری های متفاوت یادگیری عمیق برای حل این مسئله استفاده شده است. --- از معماری CNN یک بعدی در تحقیقشان استفاده کرده اند که به عنوان آزمایش دوم پیاده سازی شده است. [1] کلاس های این تحقیق شامل مثبت، منفی و خنثی می باشند در حالی که مسئله خواسته شده شامل 10 کلاس است که این تفاوت در پیاده سازی لحاظ شده است. مزیت این روش عدم وابستگی به قواعد و نگارش زبان خاص می‌باشد. لایه اول این معماری برای embedding، لایه بعدی Conv1Dو لایه آخر یک لایه fully connected layer برای dense کردن خروجی به طول مورد نظر می باشد. همچنین از اکتیویشن فانکشن softmax برای تبدیل عدد به احتمال های معنی دار و طبقه بندی نهایی استفاده شده است. (شکل 1)



شکل 1. معماری تشخیص احساسات آزمایش2

2.4. آزمایش سوم: LSTM

بسیاری از تحقیقات از معماری LSTM نیز برای حل این مسئله استفاده میکنند که به عنوان آخرین آزمایش پیاده سازی شده است. [4],[3]

3. ارزیابی

مدل	دقت روی داده train	دقت روی داده test
Chi2 + NB	-	34.8
CNN	84	28.7
LSTM	83	30

تفاوت دقت در داده آموزش و تست در هر دو مدل دیپ لرنینگ میتواند از حجم داده کم یا متفاوت بودن توزیع این دو مجموعه داده باشد.

ارزیابی با جزئیات بیشتر به صورت زیر می‌باشد:

آزمایش 2:

	precision	recall	f1-score	support
0	0.24	0.26	0.25	148
1	0.30	0.29	0.29	70
2	0.22	0.14	0.17	42
3	0.51	0.10	0.16	421
4	0.32	0.41	0.36	181
5	0.19	0.10	0.13	29
6	0.26	0.50	0.34	117
7	0.40	0.25	0.31	8
8	0.00	0.00	0.00	51
9	0.35	0.28	0.31	159
micro avg	0.30	0.23	0.26	1226
macro avg	0.28	0.23	0.23	1226
weighted avg	0.35	0.23	0.24	1226
samples avg	0.23	0.23	0.23	1226

آزمایش 3:

	precision	recall	f1-score	support
0	0.22	0.32	0.26	148
1	0.26	0.26	0.26	70
2	0.33	0.12	0.18	42
3	0.48	0.19	0.28	421
4	0.33	0.46	0.39	181
5	0.05	0.03	0.04	29
6	0.30	0.29	0.29	117
7	0.17	0.12	0.14	8
8	0.15	0.10	0.12	51
9	0.29	0.36	0.32	159
micro avg	0.31	0.27	0.29	1226
macro avg	0.26	0.23	0.23	1226
weighted avg	0.34	0.27	0.28	1226
samples avg	0.27	0.27	0.27	1226

4. آماده سازی فرمت نهایی نتایج پیشبینی

به دلیل برخی از داده ها به حهت بالا بردن دقت مدل تکرار شدند، در مرحله آخر خروجی پیشبینی شده به حالت اولیه که هر سطر میتواند به حداکثر دو کلاس تعلق داشته باشد بازمیگردد و داده های تکراری حذف میشود.

5. پیشنهادات آینده

1. استفاده از مدل ParsBert برای استخراج ویژگی از متن ها
2. اجرای مدل روی دیتاست بزرگتر
3. استفاده از معماری Bi-LSTM

6. منابع

[1] Attia, Mohammed, et al. "Multilingual multi-class sentiment classification using convolutional neural networks." *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[2] Krishnan, Hema, M. Sudheep Elayidom, and T. Santhanakrishnan. "Emotion detection of tweets using naive bayes classifier." Emotion (2017)

[3] Dashtipour, Kia, et al. "Sentiment analysis of persian movie reviews using deep learning." Entropy 23.5 (2021): 596.

[4] Su, Ming-Hsiang, et al. "LSTM-based text emotion recognition using semantic and emotional word vectors." 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia). IEEE, 2018.