گزارش تمرین سری چهارم درس بازیابی پیشرفته‌ی اطلاعات
سامانه‌ی دسته‌بند برای شبکه‌های اجتماعی

مصطفی اوجاقی۹۷۱۰۵۷۸۲، دانیال عرفانیان۹۷۱۱۰۱۵۵، کیان باختری۹۷۱۱۰۰۲۵

استاد درس: دکتر احسان‌الدین عسگری

۱۰ تیر ۱۴۰۱

چکیده: در این تمرین مجموعه‌ی داده‌هایی از شبکه‌ی اجتماعی توییتر مورد تحلیل و بررسی قرار گرفته و سیستم دسته‌بند برای توییت‌ها پیاده‌سازی شده است. کدهای این تمرین، در کنار این گزارش، در یک ژوپیتر نوت‌بوک با نام HW4 موجود است.

۱ جمع‌آوری داده

دیتاستی در این تمرین استفاده شده است که در صورت تمرین پیشنهاد شده بود. یعنی دیتاست sentiment۱۴۰ که حاوی یک میلیون و ششصد هزار توییت است که بر اساس مثبت یا منفی بودن‌شان برچسب خورده‌اند. البته با توجه به حجم بالای توییت‌های دیتاست از کسری از آن‌ها در این تمرین استفاده شده است.

۲ پیش‌پردازش

توییت‌های جمع‌آوری شده حاوی زبان‌های مختلف، اموجی، کلمات قصار، هشتگ، منشن و بسیاری از موارد دیگر بودند که با استفاده از پیش‌پردازش یا حذف شدند و یا به فرمت قابل قبولی تبدیل شدند. در خط لوله‌ی پیش‌پردازش (preprocessing pipeline) به تربیت از موارد زیر استفاده شده است:

- **حذف توییت‌های با زبان غیر انگلیسی**
- **حذف توییت‌های تکراری**
- **حذف خطوط اضافه و فاصله‌های سفید طولانی**
- **حذف منشن‌ها و هشتگ‌ها**
- **حذف تگ‌های html**
- **حذف علائم نگارشی**
- **حذف لینک‌ها و هایپرلینک‌ها**
- **حذف حروف لهجه‌دار**
- **تبدیل حروف بزرگ به کوچک**
- **کاهش حروف مکرر (پیاپی) به یک حرف**
- **گسترش عبارات فشرده**
- **حذف کلمات ایست**
- **تصحیح املای کلمات**

پس از اجرای این مراحل، با استفاده از کتابخانه‌ی nltk این توییت‌ها lemmatize و tokenize شدند تا آماده‌ی نهفته‌سازی (embedding) شوند.

۳ سامانه‌ی دسته‌بند کلاسیک

این بخش در دو مرحله انجام شد. مرحله‌ی اول درست شدن امبدینگ از توییت‌ها و مرحله‌ی دوم دسته‌بندی توییت‌ها با استفاده از بازنمایی‌هایی که از توییت‌ها موجود است. برای ایجاد بازنمایی عددی از توییت‌ها از مدل از پیش ترین شده‌ی SentenceTransformer استفاده شد. این مدل هر توییت را دریافت می‌کند و یک بردار ۳۸۴ بعدی به ازای آن توییت ایجاد می‌کند که بازنمایی آن توییت می‌باشد.

در مرحله‌ی بعد، به ازای هر توییت یک بردار عددی داریم و یک برچسب و از این جا به بعد یک مسئله‌ی یادگیری ماشین ساده داریم. داده‌ها به دسته‌های ترین و تست تقسیم شدند و توسط سه دسته‌بند معروف یعنی RandomForest ، SVM و LogisticRegression دسته‌بندی و تست شدند.

هر کدام از این مدل‌ها به دقت‌هایی در حدود ۷۰ الی ۷۳ درصد دست یافتند. بهترین نتیجه توسط مدل LogisticRegression به دست آمد که F1-score در حدود ۷۳ درصد به دست داد.

۴ سامانه‌ی دسته‌بند ترنسفورمری

برای این بخش از مدل معروف BERT استفاده شد. شیوه‌ی ترنسفر لرنینگ در این بخش به کار گرفته شد به این صورت که این مدل از پیش ترین شده است ولی باز هر داده‌های ما به سه دسته‌ی ترین، ولیدیشن و تست تقسیم شدند و مدل با داده‌ی ترین ما مقداری ترین شد یا به عبارت بهتر تنظیم (fine tune) شد. در نهایت مدل BERT توانست به دقت ۷۶ درصد دست پیدا کند.

۵ کدها

تمامی کدها و نتایج این تمرین در ژوپیتر نوت‌بوکی به نام HW4 که در کنار این گزارش است موجود هستند.