# گزارش دسته‌بندی برای شناسایی مقالات مرتبط

انجام شده توسط کیان باختری به عنوان تسک آزمایشی

۱۸ مرداد ۱۴۰۱

## ۱ تحلیل اکتشافی داده و پیش‌پردازش

در این بخش به تحلیل اکتشافی داده [۱] یعنی بررسی مجموعه‌ی داده و ویژگی‌های آن پرداخته می‌شود.

### ۱-۱ مجموعه‌ی داده

مجموعه‌ی داده‌ی مورد بررسی، یک مجموعه از مقالات [۲] با موضوعات متفاوت اما عمدتاً اخبار سیاسی و اجتماعی است که در میان آن‌ها اخبار و مقالاتی با موضوعاتی مرتبط به شبکه‌های اجتماعی و شرکت‌های بزرگ حوزه‌ی تکنولوژی وجود دارند. هدف از این بررسی ساخت یک سامانه‌ی دسته‌بند است که بتواند این مقالات را شناسایی کند.

این مجموعه‌ی داده با فرمت csv در کنار این گزارش قرار دارد که شامل ۳۳۷۶ سطر (مقاله) و ۲۱ ستون (ویژگی) می‌باشد. از میان این ۲۱ ویژگی، حدود ۷ ویژگی مربوط به زمان، مکان، و نوع مقاله است و حدود ۸ ویژگی هم به موضوع، تیتر، و متن مقاله مربوط می‌شوند. سایر ویژگی‌ها نیز معانی متفرقه‌ای دارند.

### ۱-۲ مقادیر گم شده (نامعلوم)

از میان ۳۳۷۶ مقاله‌ای که در این مجموعه لیست شده‌اند، بیشتر از ۲۵ درصدشان ویژگی اصلی یعنی خود متن مقاله را ندارند و به جای آن مقدار na یافت می‌شود. به غیر از متن اصلی که در میان ستون‌ها با عبارت text مشخص می‌شود، دو ستون دیگر هستند که بخش‌هایی از متن اصلی را در خود دارند یعنی text lead و text body که به ترتیب شامل مقدمه‌ی مقاله (مثلاً پاراگراف اول) و بدنه‌ی مقاله هستند. برای آن نمونه‌هایی که متن اصلی na است، در دو ستون دیگر هم مقادیر na هستند و این یعنی در این نمونه‌ها جایگزینی برای متن اصلی وجود ندارد و باید از مجموعه‌ی داده حذف شوند.

بعضی از ستون‌ها هم مقادیر نامعلوم زیادی دارند مانند title h2 و articleHead که چون نسبت به متن اصلی اطلاعات زیادی در بر ندارند، می‌توانند از مجموعه حذف شوند.

[۱] Exploratory data analysis

[۲] Article

### ۱-۳ ویژگی‌ها

اولین ویژگی این مجموعه‌ی داده id است که یک عدد یکتا به هر مقاله اختصاص داده شده است. سپس چند ویژگی زمانی و مکانی در مجموعه آمده است و بعد از آن‌ها شش ویژگی اصلی مرتبط با محتوای مقالات آمده‌اند.

با توجه به این که هدف اصلی در این پروژه دسته‌بندی مقالات است، می‌توان از پیش حدس زد که این دسته‌بندی بر اساس محتوای مقالات انجام می‌شود و سایر ویژگی‌ها مانند زمان انتشار و یا ژورنال خبری منتشرکننده تاثیر چندانی بر نتیجه‌ی دسته‌بندی نخواهند داشت. در ابتدا از این ویژگی‌های متفرقه صرف نظر شده و بر محتوای مقالات تمرکز شده است. در ادامه با بررسی نتایج دسته‌بندی، مشخص می‌شود که اضافه شدن این ویژگی‌ها آیا کمکی به افزایش دقت دسته‌بندی می‌کنند یا خیر. پس ویژگی‌هایی که مستقیماً با محتوای مقالات در ارتباط نیستند به صورت موقت کنار گذاشته می‌شوند. ویژگی‌های مرتبط با محتوای مقاله شامل : title h1 ، text 200 ، title h2 ، articleHead ، text ، text lead و text body هستند. برای بررسی‌های بیشتر نیازمند پیش‌پردازش هستیم.

### ۱-۴ پیش‌پردازش

مقالات حاوی کلمات قصار، هشتگ، منشن، لینک، تگ‌های html و بسیاری از موارد دیگر هستند که با استفاده از پیش‌پردازش یا حذف شدند و یا به فرمت قابل قبولی تبدیل شدند. در خط لوله‌ی پیش‌پردازش [۳] به تربیت از موارد زیر استفاده شده است:

- **حذف مقالات تکراری (در صورت وجود)**
- **حذف خطوط اضافه و فاصله‌های سفید طولانی**
- **حذف منشن‌ها و هشتگ‌ها**
- **حذف تگ‌های html**
- **حذف علائم نگارشی**

[۳] Preprocessing pipeline

- **حذف لینک‌ها و هایپرلینک‌ها**
- **تبدیل حروف لهجه‌دار**
- **تبدیل حروف بزرگ به کوچک**
- **کاهش حروف مکرر (پیاپی) به یک حرف**
- **گسترش عبارات فشرده**
- **حذف کلمات ایست**
- **تصحیح املای کلمات**

### ۱-۵ بررسی ویژگی‌ها پس از پیش‌پردازش

پیش‌پردازش متنی بر روی اطلاعات ستون‌ها اجرا شده و کلمات به فرم ساده و ریشه‌ای خود برگردانده شده‌اند تا تحلیل و بررسی آن‌ها ممکن شود. شایان ذکر است که این پیش‌پردازش باعث کوتاه‌تر شدن طول عبارات و متن‌ها می‌شود چرا که عبارات غیر مهم مانند کلمات ایست [۴] از متن‌ها حذف می‌شوند.

ستون title h2 بیش از ۹۰ درصد مقادیرش نامعلوم بوده و حذف شد. ستون text 200 حاوی ۲۰۰ کاراکتر اول هر مقاله است که اطلاعات جدیدی در بر ندارد و حذف می‌شود. ستون title h1 حاوی تیتر مقالات است. در شکل‌های ۱ و ۲ به ترتیب هیستوگرام کلمات پرتکرار در ستون title h1 و توزیع تعداد کلمات در این ستون نمایش داده شده‌اند. همچنین ستون articleHead حاوی عبارات اصلی مرتبط با موضوع مقالات است. در شکل‌های ۳ و ۴ به ترتیب هیستوگرام کلمات پرتکرار در ستون articleHead و توزیع تعداد کلمات در این ستون نمایش داده شده‌اند. همانطور که در شکل ۴ پیدا است، این ستون اطلاعات اندکی در بر دارد و می‌توان این ستون را کنار گذاشت.

ستون text lead مقدمه یا پاراگراف ابتدایی مقالات را در بر دارد. این موضوع از آن جا پیدا است که هم می‌توان تعداد کمی از نمونه‌ها را به صورت دستی چک کرد و هم می‌توان از مقایسه‌ی متون و امتیاز دهی برای تمامی نمونه‌ها بهره برد. شکل ۵ توزیع امتیاز تشابه میان محتوای ستون text lead و همان مقدار از ابتدای محتوای ستون text را نشان می‌دهد. این نمودار همچنین همین نوع امتیازدهی برای تشابه میان محتوای ستون text body و ادامه‌ی محتوای ستون text را در بر دارد. برای امتیازدهی از معیار fuzz ratio استفاده شده است. این نمودار به سادگی نمایشگر آن است که ستون‌های text lead و text body اطلاعات اضافه‌تری از متن اصلی مقاله در بر ندارند و برای ساخت دسته‌بند اولیه می‌توان فقط متن اصلی را در نظر گرفت. به همین دلیل است که در انتهای نوت بوک EDA یک دیتافریم به نام simple df درست شده و برای ادامه‌ی کار از آن استفاده شده است.

ستون text حاوی متن اصلی هر مقاله است. پس از پیش‌پردازش تعداد کلمات مهم هر کدام از این مقالات در حدود ۵۰۰ کلمه بوده است (شکل ۶). مقالاتی که برچسب «مرتبط» داشته‌اند کلمات پرتکرارشان در شکل ۷ و همچنین کلمات پرتکرار مقالات با برچسب «نامربوط» در شکل ۸ قابل مشاهده است. از این نمودارها پیدا است که مربوط یا نامربوط بودن مقالات، به شبکه‌های اجتماعی در فضاهای مجازی و شرکت‌های بزرگ حوزه‌ی تکنولوژی ارتباط دارد.

[۴] stop words

## ۲ افزایش داده

نظر به متعادل نبودن تعداد داده‌های کلاس‌ها، برای ترین کردن مدل‌ها هم از وزن‌دهی بیشتر به کلاس با تعداد کم‌تر استفاده شد و هم از افزایش داده (data augmentation).

شیوه‌ی افزایش داده به ساده‌ترین صورت ممکن اینگونه انجام شد که برای کلاس با تعداد داده‌ی کم‌تر، یک توزیع روی کلمات آن کلاس و همچین تعداد کلمات به ازای هر ویژگی به دست آمد و برای درست کردن داده‌ی جدید از آن توزیع‌ها نمونه گرفته شد. بدین صورت تعداد داده‌های کلاس‌ها متعادل شد.

البته که نتایج دسته‌بندی هم با افزایش داده و هم بدون آن گزارش شده‌اند. از این نتایج پیدا است که افزایش داده نقش مثبتی در بهبود دقت داشته است.

## ۳ دسته‌بندی و مهندسی ویژگی‌ها

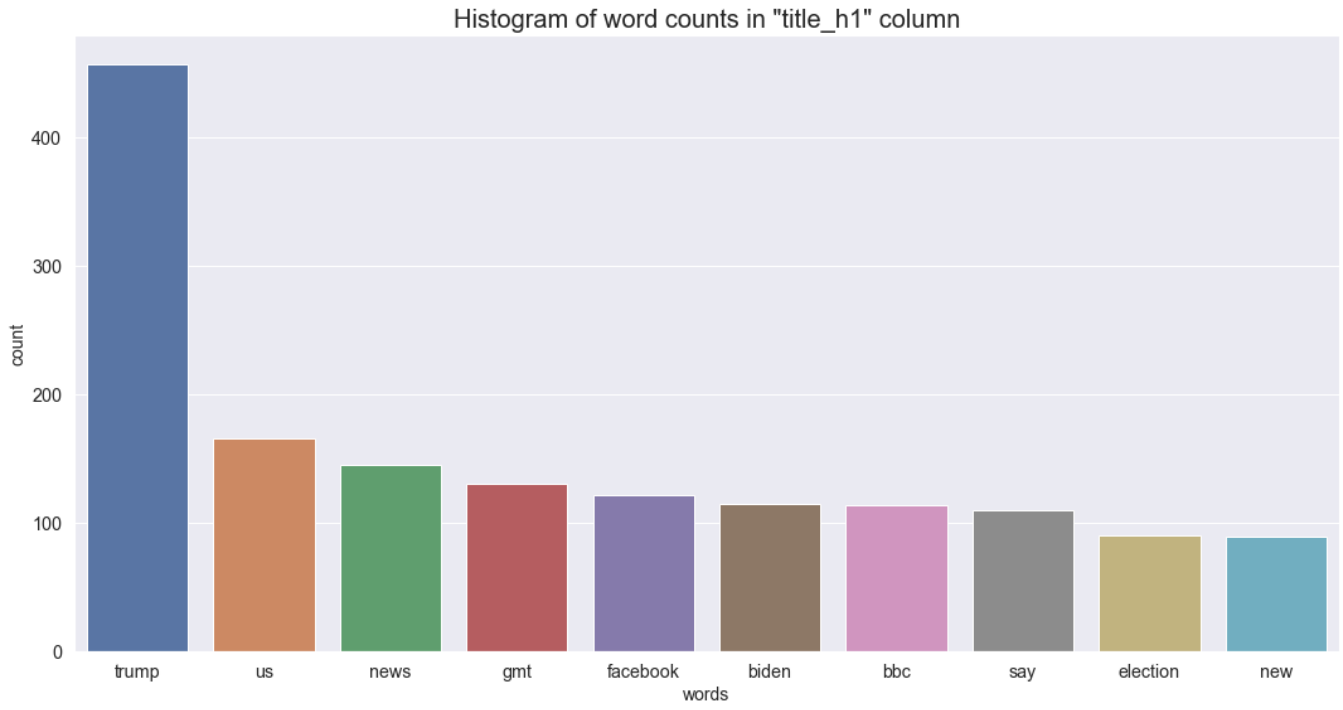
برای انجام دسته‌بندی ترکیب‌های متنوعی از ویژگی‌ها انتخاب شدند تا مورد آزمود قرار گیرند:

- **عنوان مقاله**
- **عنوان مقاله و جمله‌ی اول**
- **عنوان مقاله و پنج جمله‌ی اول**
- **عنوان مقاله و ده جمله‌ی اول**
- **عنوان مقاله و پاراگراف اول**
- **عنوان مقاله و جمله‌ی اول از هر پاراگراف**

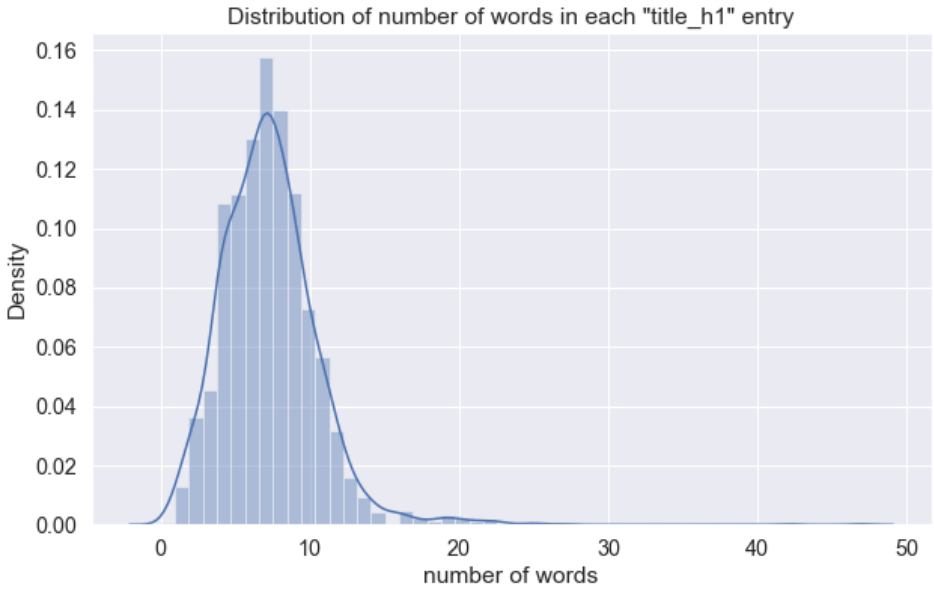
به ازای هر کدام از این ویژگی‌ها، دسته‌بندی با دو روش اصلی صورت گرفت:

- **امبدینگ توسط مدل عمیق و دسته بندی با مدل کلاسیک**
- **دسته‌بندی با مدل عمیق**

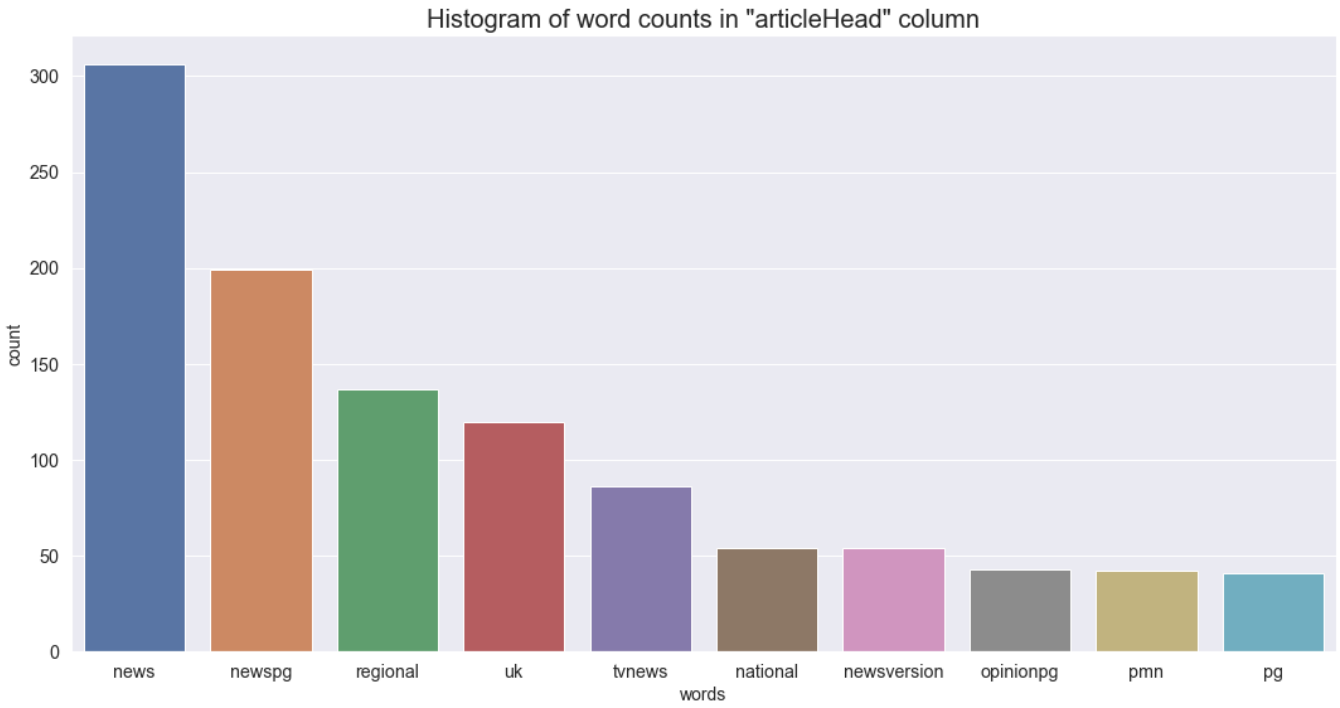
نتایج تمامی دسته‌بندی‌ها در شکل ۹ قابل مشاهده هستند. بیان این نکته ضروری است که به عقیده‌ی نگارنده، این نتایج نمی‌توانند پایان این مسئله در حالت واقعی باشند و اگر این مسئله واقعی و زنده بود، باید زمان بیشتری صرف می‌شد تا نتایج بهبود یابند. به عقیده‌ی نگارنده، این گزارش و نتایج آن صرفاً حالت یک نسخه‌ی نمایشی از پروسه‌ی حقیقی را دارند.



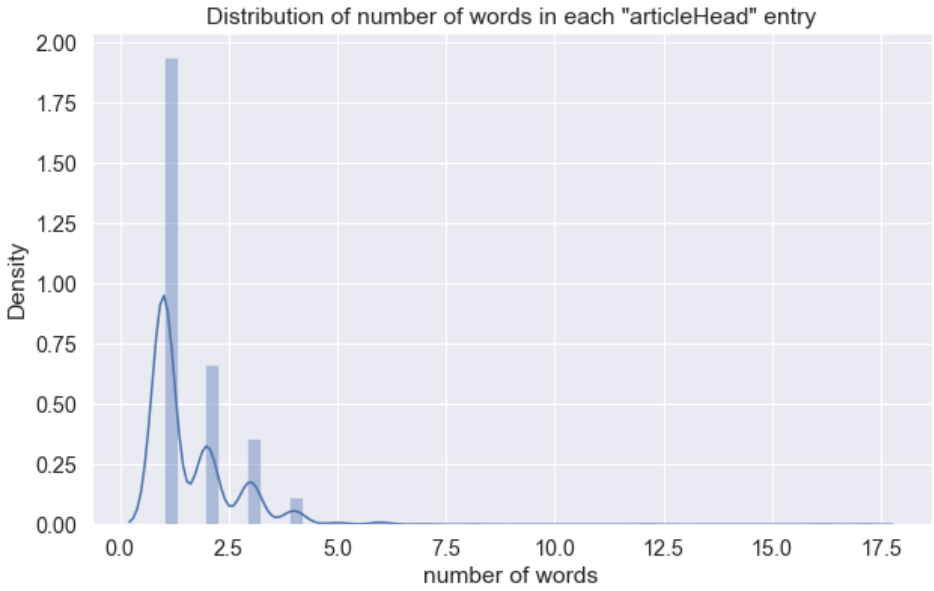

شکل ۱:  نمودار ستونی کلمات پر بسامد در ستون title h1



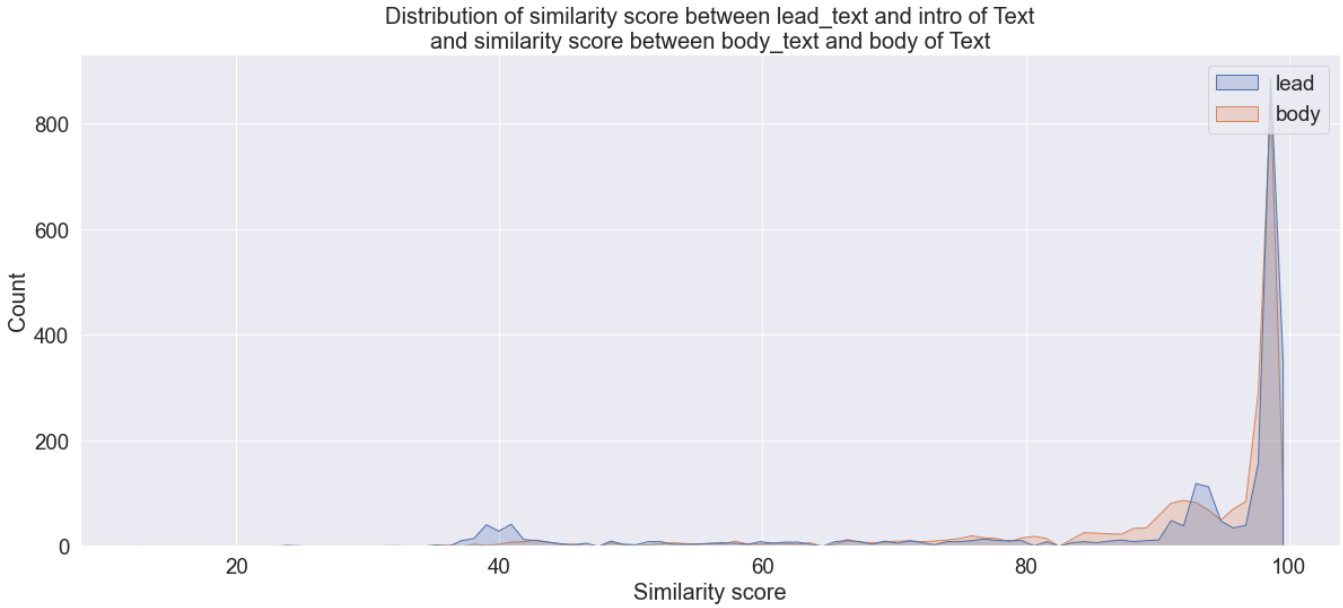

شکل ۲:  توزیع تعداد کلمات در ستون title h1



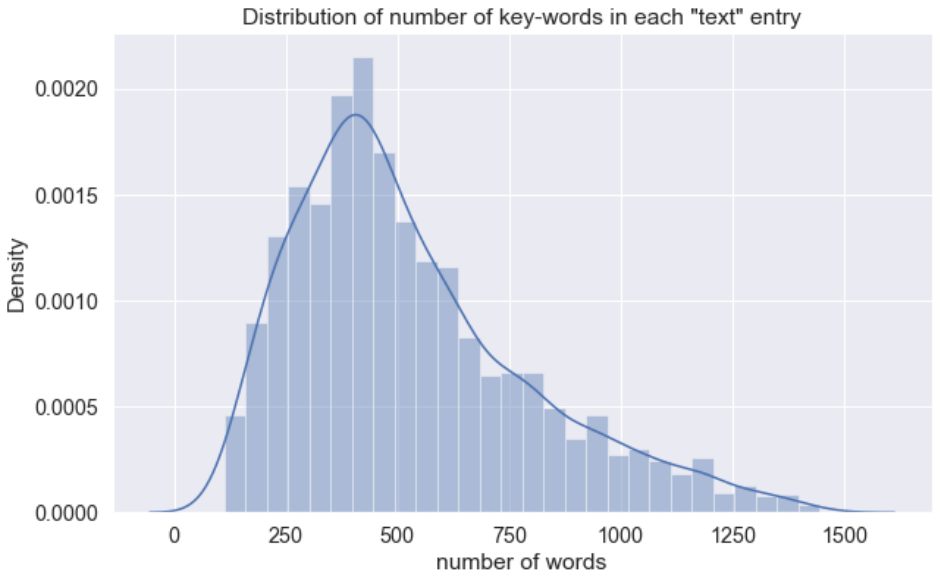

شکل ۳: نمودار ستونی کلمات پر بسامد در ستون articleHead



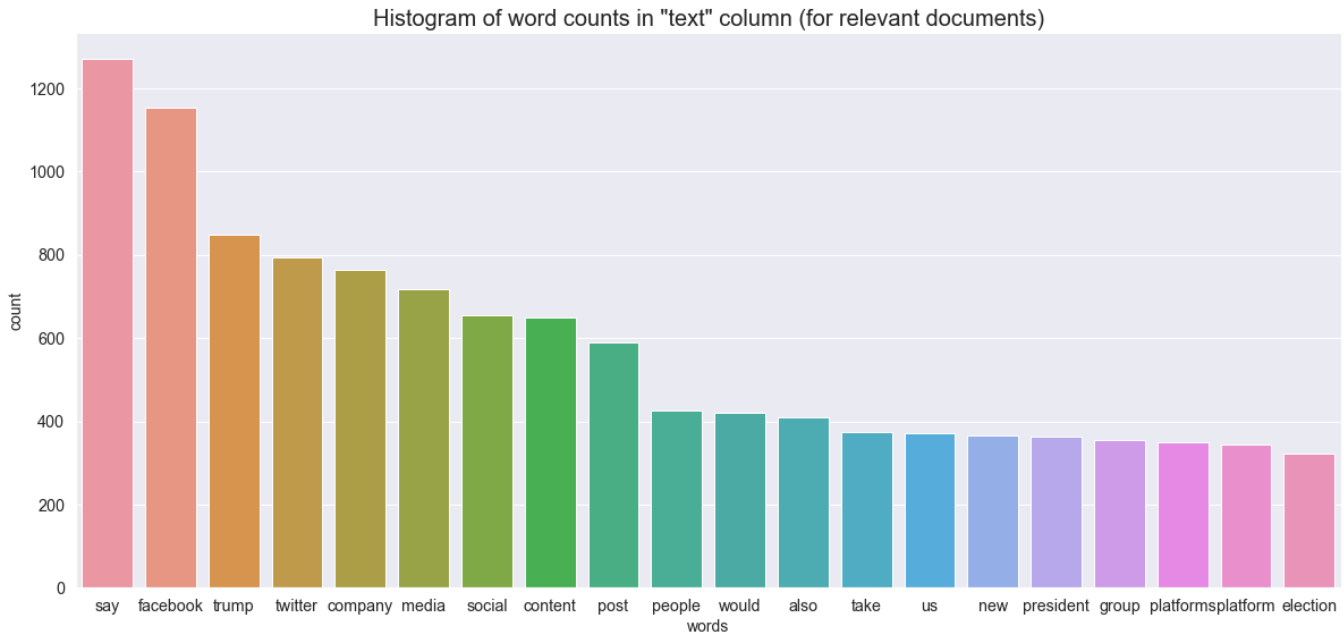

شکل ۴: توزیع تعداد کلمات در ستون articleHead



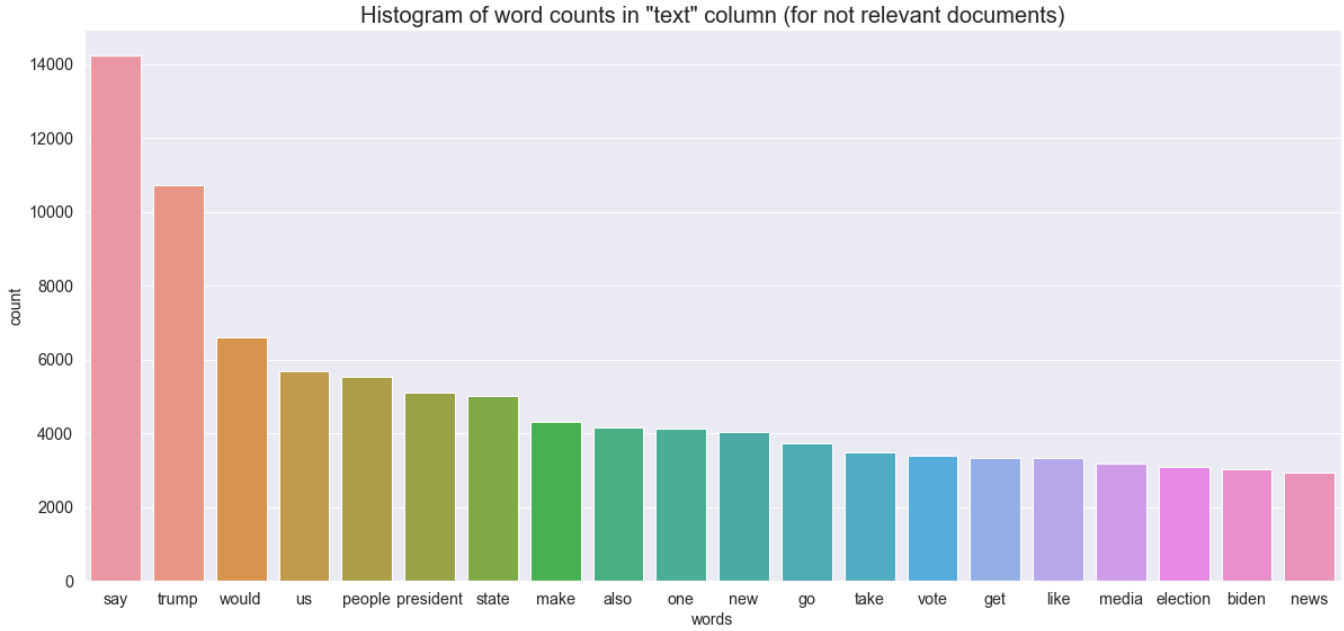

شکل ۵: توزیع امتیاز تشابه میان محتوای ستون text lead و همان مقدار از ابتدای ستون text برای هر نمونه




شکل ۶: توزیع تعداد کلمات مهم در داده‌های ستون text (متن اصلی مقاله)




شکل ۷:  هیستوگرام کلمات پرتکرار در مقالاتی که برچسب مربوط دارند




شکل ۸:  هیستوگرام کلمات پرتکرار در مقالاتی که برچسب نامربوط دارند

### ۳-۱ امبدینگ توسط مدل عمیق و دسته‌بندی با مدل‌های کلاسیک

برای امبدینگ از کتابخانه‌ی SentenceTransformer استفاده شد که مدل‌های از پیش ترین شده‌ی زیادی را در بر دارد. برای این بخش، از دو مدل عمیق برای تولید امبدینگ‌ها استفاده شد. مدل all-MiniLM-L6-v2 که هر متن ورودی‌اش را به یک بردار ۳۸۴ بعدی می‌نگارد و مدل all-distilroberta-v1 که هر متن ورودی‌اش به برداری ۷۶۸ بعدی نگاشته می‌شود.

برای دسته‌بندی امبدینگ‌ها، مدل‌های زیر مورد استفاده قرار گرفتند:

- **XGBoost**
- **SVM**
- **LogisticRegression**
- **RandomForest**

نتایج دسته بندی در شکل ۹ به نمایش درآمده است. نکته‌ای که در این نتایج کمی تعجب‌برانگیز است، بهتر نبودن نتایج مدل all-distilroberta-v1 است که حدس بنده آن است که شاید مشکل کوچکی در امبدینگ وجود دارد. بدیهی است که اگر درگیر یک مسئله‌ی واقعی بودیم زمان بیشتری صرف حصول اطمینان از نتایج می‌شد.

### ۳-۲ دسته‌بندی با مدل عمیق

در این بخش، مدل DistilBERT و البته حالت خاص آن برای دسته‌بندی یعنی DistilBertForSequenceClassification استفده شد. حالت از پیش ترین شده‌ی این مدل از کتابخانه‌ی transformers دانلود شده و روی دیتاست این مسئله مجدداً آموزش داده شده و یا به عبارتی تنظیم شده است. برای این کار از میان ۱۰۵ لایه‌ی این مدل، ۹۵ لایه‌ی اولیه فریز شده و فقط ۱۰ لایه‌ی آخرآزاد گذاشته شدند.

با توجه به نتایجی که از دسته‌بندی با مدل‌های کلاسیک به دست آمده بود، این مدل دیگر برای تمامی حالت ویژگی‌ها آزموده نشد و فقط برای ویژگی‌های عنوان و عنوان به علاوه‌ی ده جمله‌ی اول و عنوان به علاوه‌ی پاراگراف اول مورد آزمون قرار گرفت.

البته که در یک مسئله‌ی صعنتی تمامی حالات تست می‌شوند اما نگارنده ناچار به مدیریت زمان بوده و امکان ترین کردن مدل برای تمامی حالت وجود نداشت. هرچند که همین سه حالت هم به صورت کامل ترین نشدند و هر کدام در حدود ۱۵ اپیاک ترین شدند. علاوه بر یادگیری، تنظیم هایپرپارامترها هم نیازمند کار مفصل‌تری بود. نتایج این قسمت در یک عدد خلاصه می‌شود که در شکل ۹ قابل مشاهده است. این نتیجه در نظر نگارنده نتیجه‌ی نهایی این مدل نیست و با صرف زمان بیشتر برای تنظیم هایپرپارامترها و البته زمان بیشتر برای ترین شدن مدل، می‌شود این عدد را به شکل قابل توجهی بهبود داد.

## ۴ خوشه‌بندی

داده‌ها با استفاده از الگوریتم K-means خوشه‌بندی شده و سپس با استفاده از الگوریتم کاهش بعد PCA به دو بعد آورده شدند تا در نمودار قابل نمایش باشند. شکل ۱۳ نمایش این نمودار است. اگر برچسب داده‌ها برچسب واقعی آن‌ها باشند، شکل ۱۴ به دست می‌آید.

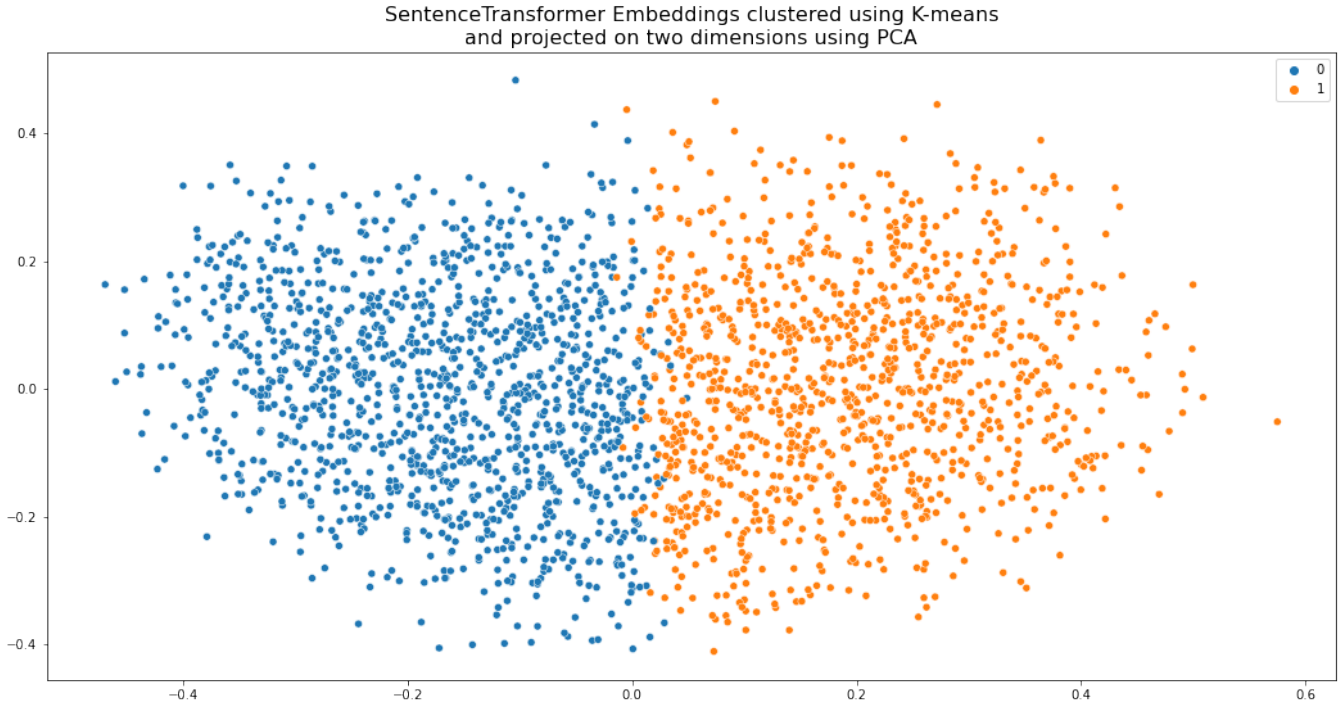
| Embedding with "all-MiniLM-L6-v2" model: each doc goes to a 384 dim vector<br>Classification with classic algorithms (without data augmentation) | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Title | | | | Title + First Sentence | | | | Title + 5 first sentences | | | | Title + 10 first sentences | | | | Title + First Paragraph | | | | Title + First sentence of each paragraph | | | |
| | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc |
| XGBoost | 0.92 | 0.46 | 0.35 | 0.66 | 0.92 | 0.41 | 0.3 | 0.67 | 0.93 | 0.53 | 0.44 | 0.67 | 0.92 | 0.47 | 0.37 | 0.62 | 0.92 | 0.51 | 0.43 | 0.62 | 0.91 | 0.38 | 0.3 | 0.53 |
| SVM | 0.92 | 0.6 | 0.69 | 0.54 | 0.92 | 0.6 | 0.67 | 0.55 | 0.93 | 0.67 | 0.8 | 0.57 | 0.92 | 0.63 | 0.76 | 0.54 | 0.92 | 0.62 | 0.72 | 0.55 | 0.91 | 0.59 | 0.7 | 0.51 |
| LogisticRegression | 0.9 | 0.59 | 0.76 | 0.49 | 0.89 | 0.57 | 0.8 | 0.44 | 0.89 | 0.59 | 0.83 | 0.46 | 0.9 | 0.61 | 0.85 | 0.47 | 0.89 | 0.58 | 0.81 | 0.45 | 0.89 | 0.59 | 0.83 | 0.45 |
| RandomForrest | 0.92 | 0.4 | 0.28 | 0.71 | 0.91 | 0.19 | 0.11 | 0.6 | 0.92 | 0.39 | 0.28 | 0.68 | 0.91 | 0.3 | 0.2 | 0.58 | 0.91 | 0.3 | 0.2 | 0.58 | 0.91 | 0.18 | 0.11 | 0.5 |

| Embedding with "all-MiniLM-L6-v2" model: each doc goes to a 384 dim vector<br>Classification with classic algorithms (with data augmentation) | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Title | | | | Title + First Sentence | | | | Title + 5 first sentences | | | | Title + 10 first sentences | | | | Title + First Paragraph | | | | Title + First sentence of each paragraph | | | |
| | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc |
| XGBoost | 0.92 | 0.55 | 0.5 | 0.6 | 0.93 | 0.49 | 0.39 | 0.66 | 0.92 | 0.54 | 0.52 | 0.56 | 0.92 | 0.59 | 0.65 | 0.55 | 0.93 | 0.59 | 0.54 | 0.64 | 0.92 | 0.6 | 0.65 | 0.56 |
| SVM | 0.92 | 0.57 | 0.56 | 0.59 | 0.93 | 0.58 | 0.5 | 0.69 | 0.94 | 0.66 | 0.65 | 0.67 | 0.94 | 0.65 | 0.67 | 0.64 | 0.94 | 0.65 | 0.63 | 0.67 | 0.92 | 0.57 | 0.61 | 0.53 |
| LogisticRegression | 0.93 | 0.6 | 0.59 | 0.6 | 0.93 | 0.57 | 0.52 | 0.62 | 0.93 | 0.63 | 0.67 | 0.6 | 0.93 | 0.65 | 0.76 | 0.57 | 0.93 | 0.61 | 0.59 | 0.63 | 0.9 | 0.57 | 0.76 | 0.46 |
| RandomForrest | 0.93 | 0.54 | 0.46 | 0.66 | 0.92 | 0.26 | 0.17 | 0.64 | 0.92 | 0.43 | 0.33 | 0.62 | 0.92 | 0.54 | 0.52 | 0.57 | 0.91 | 0.3 | 0.2 | 0.55 | 0.91 | 0.42 | 0.35 | 0.53 |

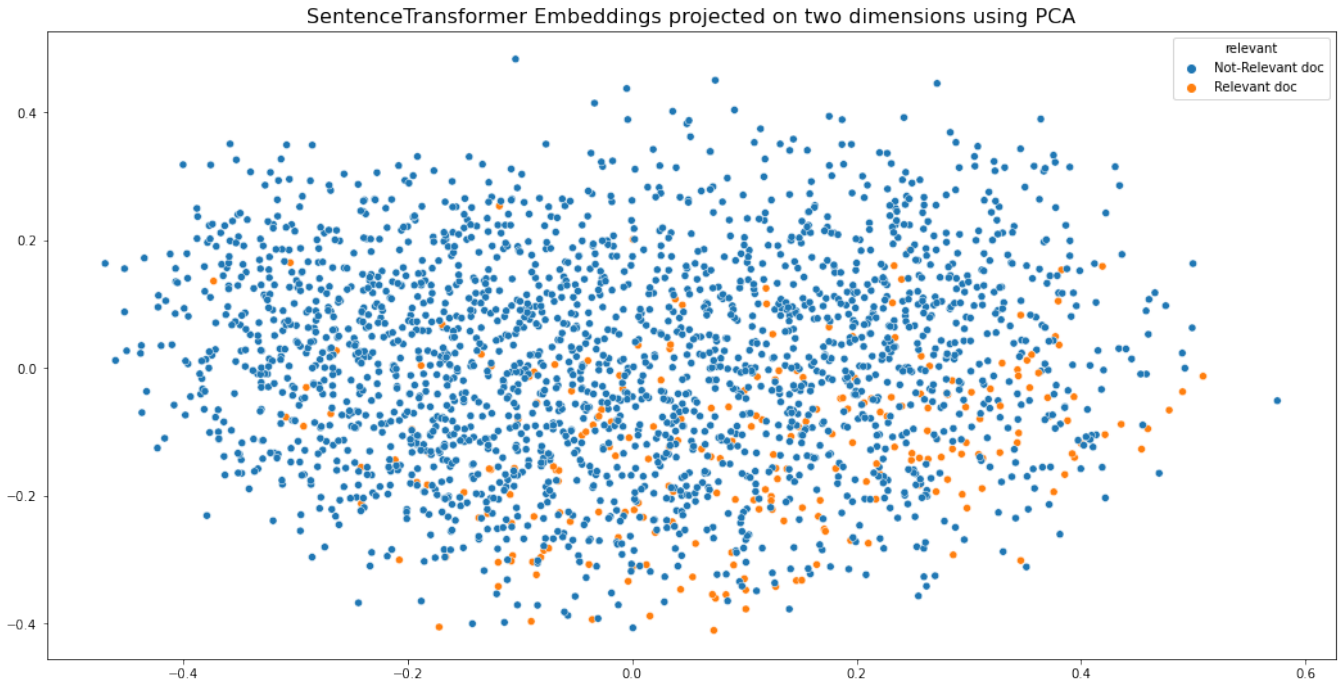
| Embedding with "all-distilroberta-v1 " model: each doc goes to a 768 dim vector<br>Classification with classic algorithms (with data augmentation) | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Title | | | | Title + First Sentence | | | | Title + 5 first sentences | | | | Title + 10 first sentences | | | | Title + First Paragraph | | | | Title + First sentence of each paragraph | | | |
| | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc | acc | f1 | recall | perc |
| XGBoost | 0.93 | 0.54 | 0.48 | 0.62 | 0.93 | 0.53 | 0.44 | 0.67 | 0.91 | 0.41 | 0.35 | 0.5 | 0.91 | 0.42 | 0.35 | 0.51 | 0.91 | 0.4 | 0.31 | 0.55 | 0.89 | 0.39 | 0.37 | 0.41 |
| SVM | 0.93 | 0.6 | 0.57 | 0.63 | 0.95 | 0.67 | 0.61 | 0.75 | 0.93 | 0.58 | 0.52 | 0.65 | 0.93 | 0.58 | 0.5 | 0.69 | 0.94 | 0.62 | 0.57 | 0.67 | 0.91 | 0.5 | 0.48 | 0.51 |
| LogisticRegression | 0.93 | 0.59 | 0.57 | 0.61 | 0.93 | 0.59 | 0.54 | 0.66 | 0.93 | 0.53 | 0.46 | 0.62 | 0.92 | 0.49 | 0.41 | 0.61 | 0.92 | 0.49 | 0.41 | 0.61 | 0.89 | 0.49 | 0.57 | 0.43 |

| Classification with deep network: "distilbert-base-uncased" (15 epoch on last 10 layers) | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| distilbert-base-uncased | | | | | | | | | | | | | | | | | | 0.71 | | | | | | |

شکل ۹: نتایج دسته‌بندی




شکل ۱۰: خوشه‌بندی امبدینگ SentenceTransformer و کاهش بعد توسط الگوریتم PCA




شکل ۱۱: کاهش بعد امبدینگ SentenceTransformer با استفاده از الگوریتم PCA (برچسب‌های واقعی)