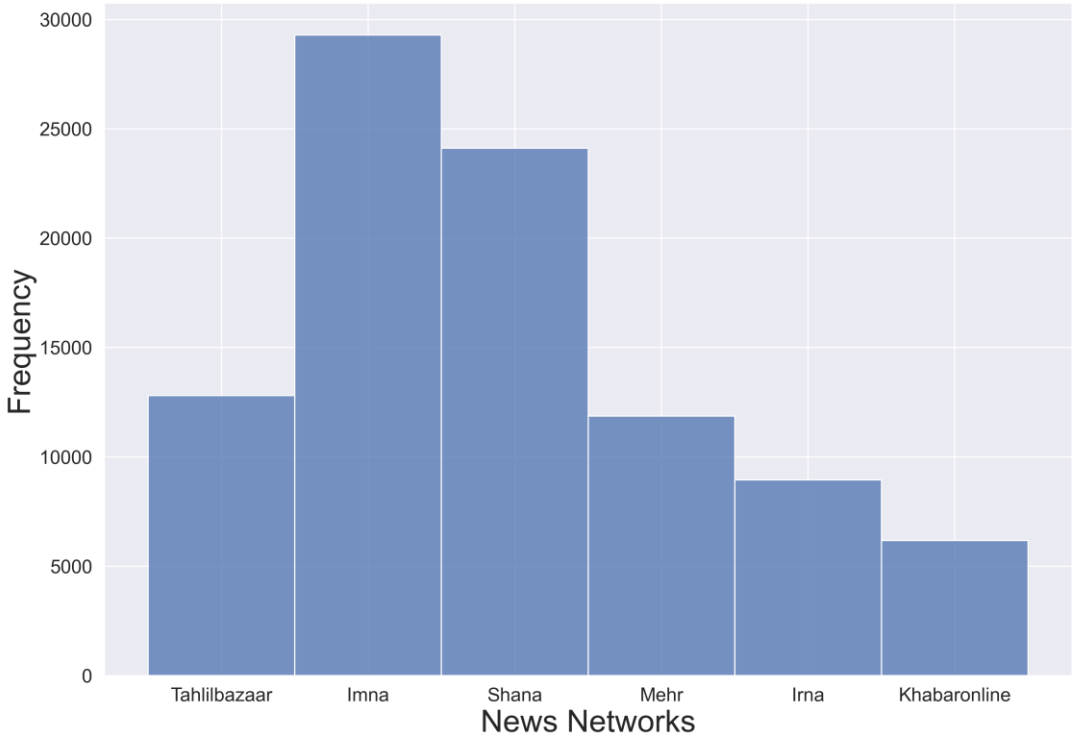
توضیحات دیتاست خلاصه‌سازی متن هوشواره

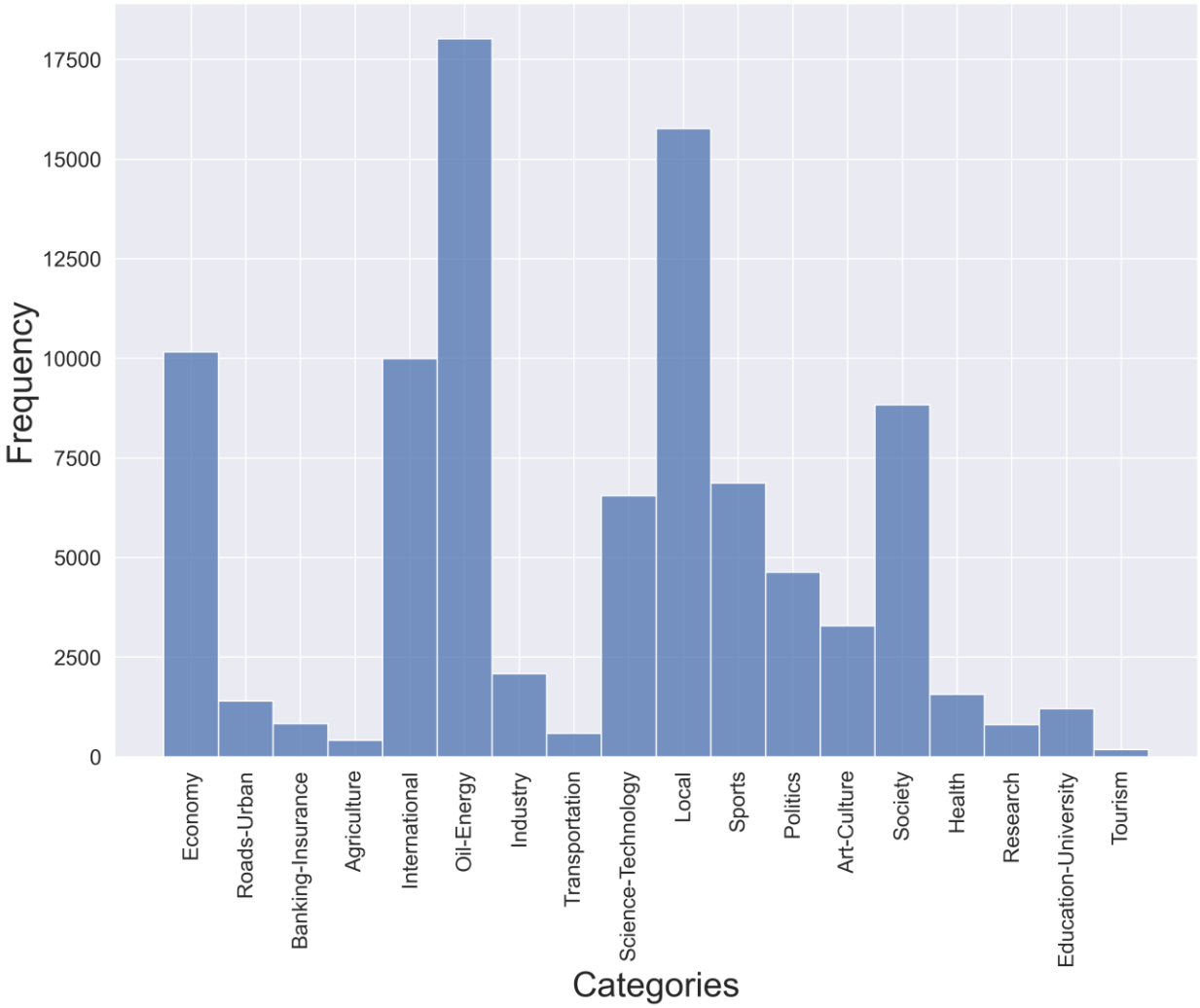
این دیتاست شامل ۹۳،۲۰۷ رکورد است که در قالب ۳ بخش داده‌ی ترین، ولیدیشن و تست در فایل‌های CSV قرار داده شده است. دیتاست حاضر شامل متن خبر، تیتر خبر، خلاصه خبر، لینک خبر و دسته‌بندی خبر در زبان فارسی و انگلیسی می‌باشد و از آن در جهت اهدافی همچون تولید متن، تولید تیتر و تشخیص دسته بندی خبر می‌تواند به کار برده شود.

این مجموعه داده از خبرگزاری‌های فارسی متفاوتی بدست آمده است که توزیع داده بدست آمده از هریک از خبرگزاری‌ها در شکل ۱ قابل مشاهده است.



شکل ۱-توزیع داده براساس خبرگزاری‌های مختلف

همچنین ۱۸ دسته‌بندی موضوعی در این مجموعه داده در نظر گرفته شده است. در میان دسته‌بندی‌ها، دسته‌بندی اخبار سوخت و انرژی بیشترین میزان رکورد و اخبار توریستی کمترین میزان خبر را به خود اختصاص داده‌اند. توزیع موضوعات اخبار موجود در مجموعه داده در شکل ۲ قابل مشاهده است.



شکل ۲-توزیع موضوع اخبار موجود در مجموعه داده

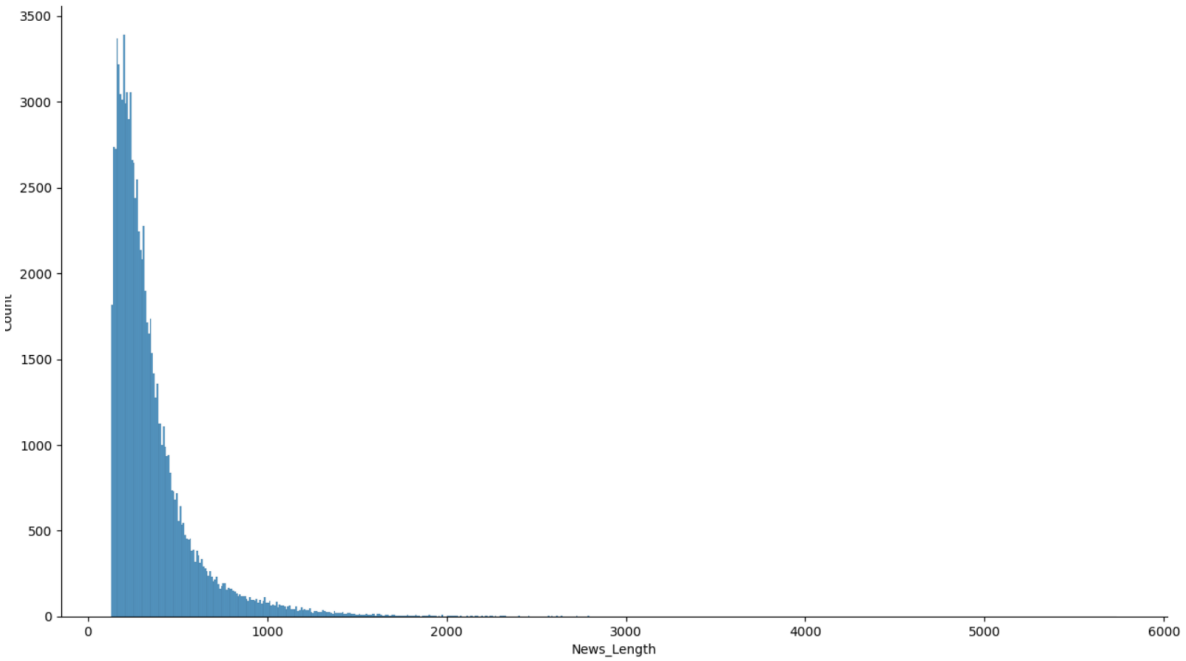
شکل ۳ ابر لغات(word cloud) مربوط به این مجموعه داده را نمایش می‌دهد.

شکل ۳-ابر لغات اخبار موجود در مجموعه داده

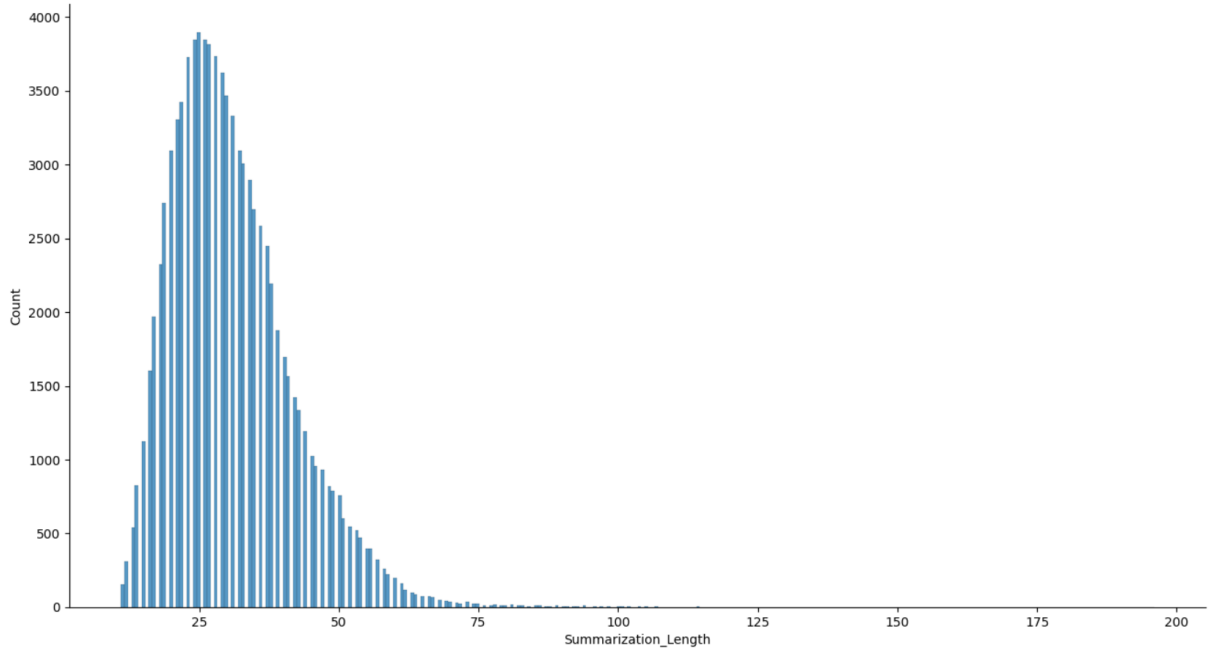
در جهت پیش‌پردازش داده، تمامی کاراکترهای مربوط به خط جدید(newline) حذف و فاصله جایگزین آن‌ها شده است. به علاوه عملیات نرمال‌سازی متن با استفاده از کتابخانه هضم[۱]صورت پذیرفته‌است. در انتها دادگان به توکن‌های کلمه‌ای تقسیم شده و تعداد توکن‌های موجود در متن هر خبر و خلاصه مربوط به آن در ستون‌های جداگانه قرار داده شده‌اند.

شکل ۴ و شکل ۵ به ترتیب توزیع تعداد توکن‌های موجود در متن و خلاصه اخبار را نمایش می‌دهند.

[۱]Hazm Library



شکل ۴-توزیع تعداد کلمات(توکن‌ها) موجود در متن اخبار



شکل ۵-توزیع تعداد کلمات(توکن‌ها) موجود در خلاصه اخبار

نمونه‌ای از ستون‌هایی از دادگان که در این پروژه مورد استفاده خواهد بود در شکل ۶ آورده شده است.

	Summarization	News	Summarization_Lengh	News_Length
0	مدیرعامل شرکت ملی نفت، عملکرد مدیریت امور الملل این شرکت را در دوران تحریم بسیاربیند هوشمندانه خواند و گفت: امور بین الملل در دوران ها نیز می‌تواند نقش بزرگی در تسریع‌پس از تحریم .روند توسعه داشته باشد	به گزارش شانا، علی کاردر امروز (۲۷ دی ماه) در مراسم تودیع محسن قمصری، مدیر سابق امور بین الملل شرکت ملی نفت ایران و معارفه سعید خوشرو، مدیر جدید امور بین الملل این الملل به عنوان یکی ازشرکت، گفت: مدیریت امور بین تاثیرگذارترین مدیریت‌های شرکت ملی نفت ایران...	39	245
1	سرپرست مدیریت برنامه‌ریزی و توسعه شرکت ملی صنایع پتروشیمی گفت: تنوع محصولات پتروشیمی ایران با بهره‌برداری از طرح‌های جهش دوم و سوم .صنعت پتروشیمی افزایش می‌یابد	به گزارش شانا به نقل از شرکت ملی صنایع پتروشیمی، علی‌اصغر گودرزی‌فراهانی با اشاره به اینکه همه طرح‌های در حال اجرای صنعت پتروشیمی براساس پیشرفت فیزیکی و پیش‌بینی زمان راه‌اندازی در قالب طرح‌های جهش دوم و سوم تقسیم‌بندی شده‌اند، اظهار کرد: انتظار داریم که ط...	28	379
2	پالایشگاه گاز خانگیران با هدف معرفی گوگرد بنتونیتی برای تولید کود مرغوب ویژه مصارف کشاورزی در .شانزدهمین نمایشگاه کشاورزی اردبیل شرکت کرد	به گزارش شانا به نقل از شرکت پالایش گاز شهید هاشمی‌نژاد، جمعی از کارشناسان این پالایشگاه با هدف معرفی محصول کود گوگرد بنتونیتی به کشاورزان در شانزدهمین نمایشگاه تخصصی کشاورزی، دام، طیور، ماشین‌آلات و صنایع وابسته در اردبیل حضور پیدا کردند و به معرفی دستاور...	23	325
3	سخنگوی شورای شهر شیراز گفت: روند عمرانی و شهرسازی در این شهر با فروکش کردن حساسیت‌های .کرونایی به حالت عادی خود باز می‌گردد	به گزارش خبرنگار ایمنا، سعید نظری در صفحه اینستاگرام خود نوشت: «در حالیکه شهر شیراز نخستین روزهای سال ۹۹ را با شیوع بیماری کرونا با حساسیت ویژه طی می‌کند، اعضای شورای اسلامی شهر شیراز در تلاشند همسو با ستاد ملی مبارزه با کرونا، مخاطرات ناشی از شیوع این ویر...	25	210

شکل ۶-چند نمونه از رکوردهای موجود در مجموعه داده نهایی(ستون News ادامه دارد و بخش ابندایی آن نمایش داده شده است).