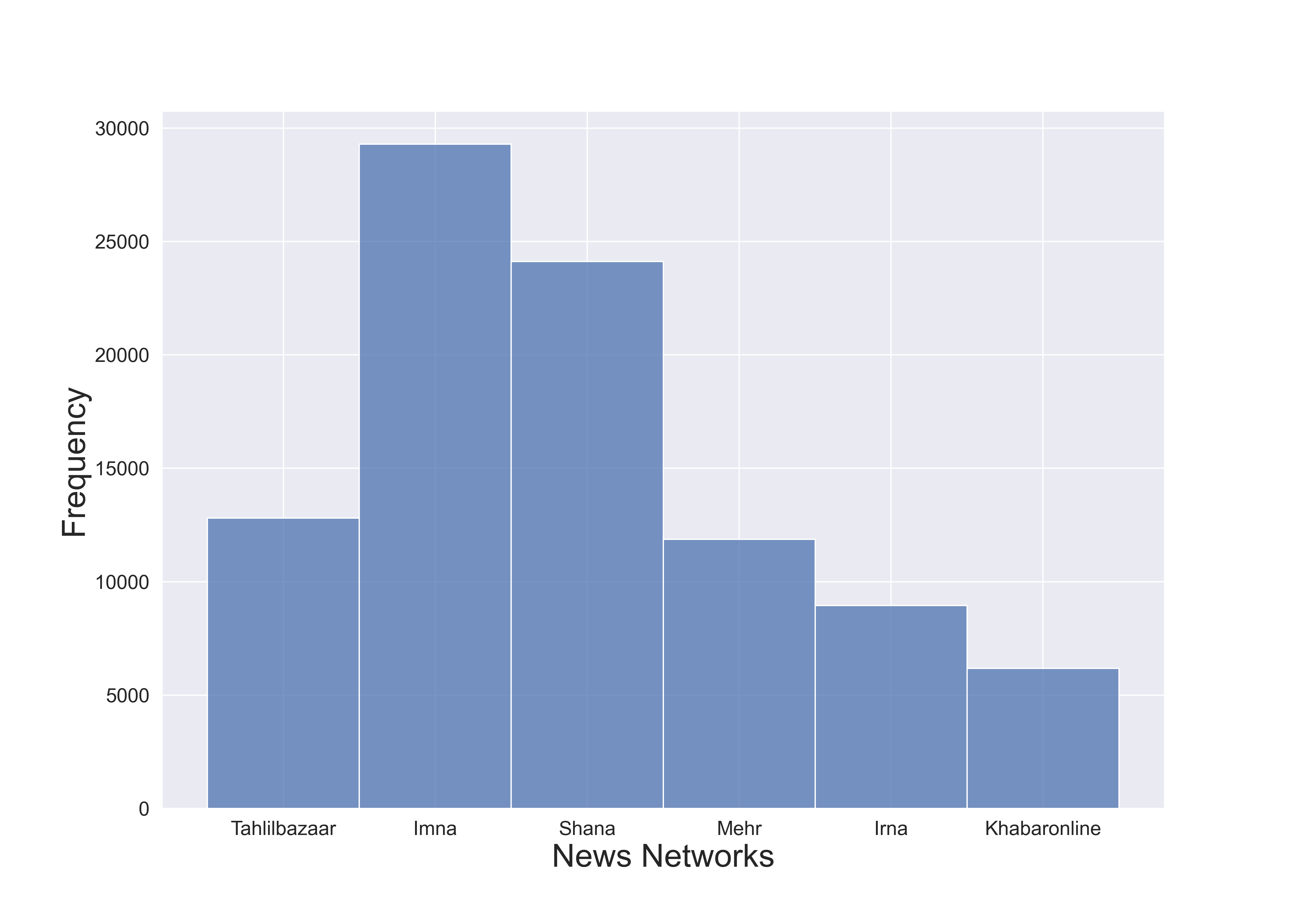
## **توضیحات دیتاست خلاصه‌سازی متن هوشواره**

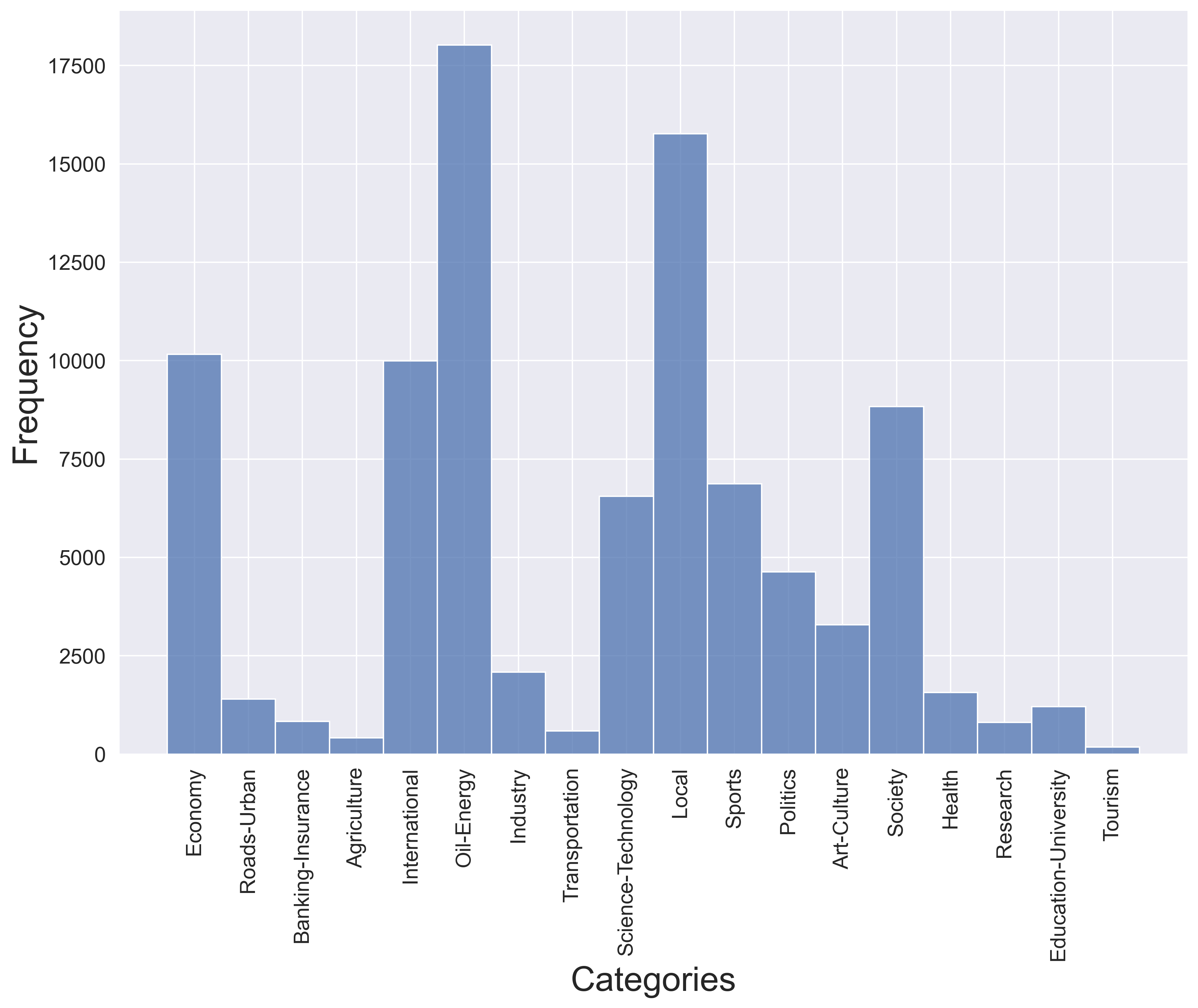
این دیتاست شامل 93،207 رکورد است که در قالب 3 بخش داده‌ی ترین، ولیدیشن و تست در فایل‌های CSV قرار داده شده است. دیتاست حاضر شامل متن خبر، تیتر خبر، خلاصه خبر، لینک خبر و دسته‌بندی خبر در زبان فارسی و انگلیسی می‌باشد و از آن در جهت اهدافی همچون تولید متن، تولید تیتر و تشخیص دسته بندی خبر می‌تواند به کار برده شود.

این مجموعه داده از خبرگزاری‌های فارسی متفاوتی بدست آمده است که توزیع داده بدست آمده از هریک از خبرگزاری‌ها در شکل 1 قابل مشاهده است.



**شکل 1-توزیع داده براساس خبرگزاری‌های مختلف**

همچنین 18 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 2 قابل مشاهده است.



**شکل 2-توزیع موضوع اخبار موجود در مجموعه داده**

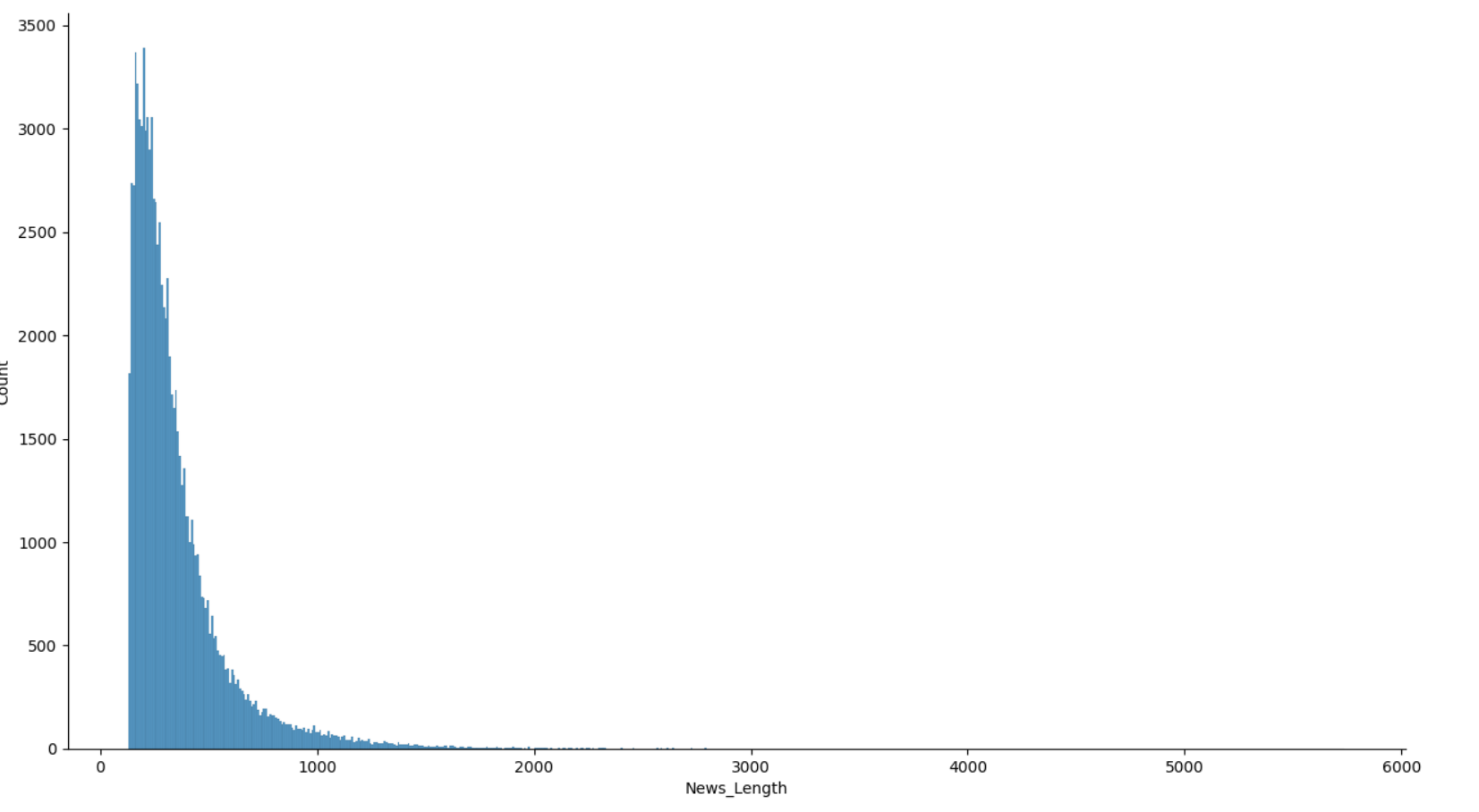
شکل 3 ابر لغات(word cloud) مربوط به این مجموعه داده را نمایش می‌دهد.



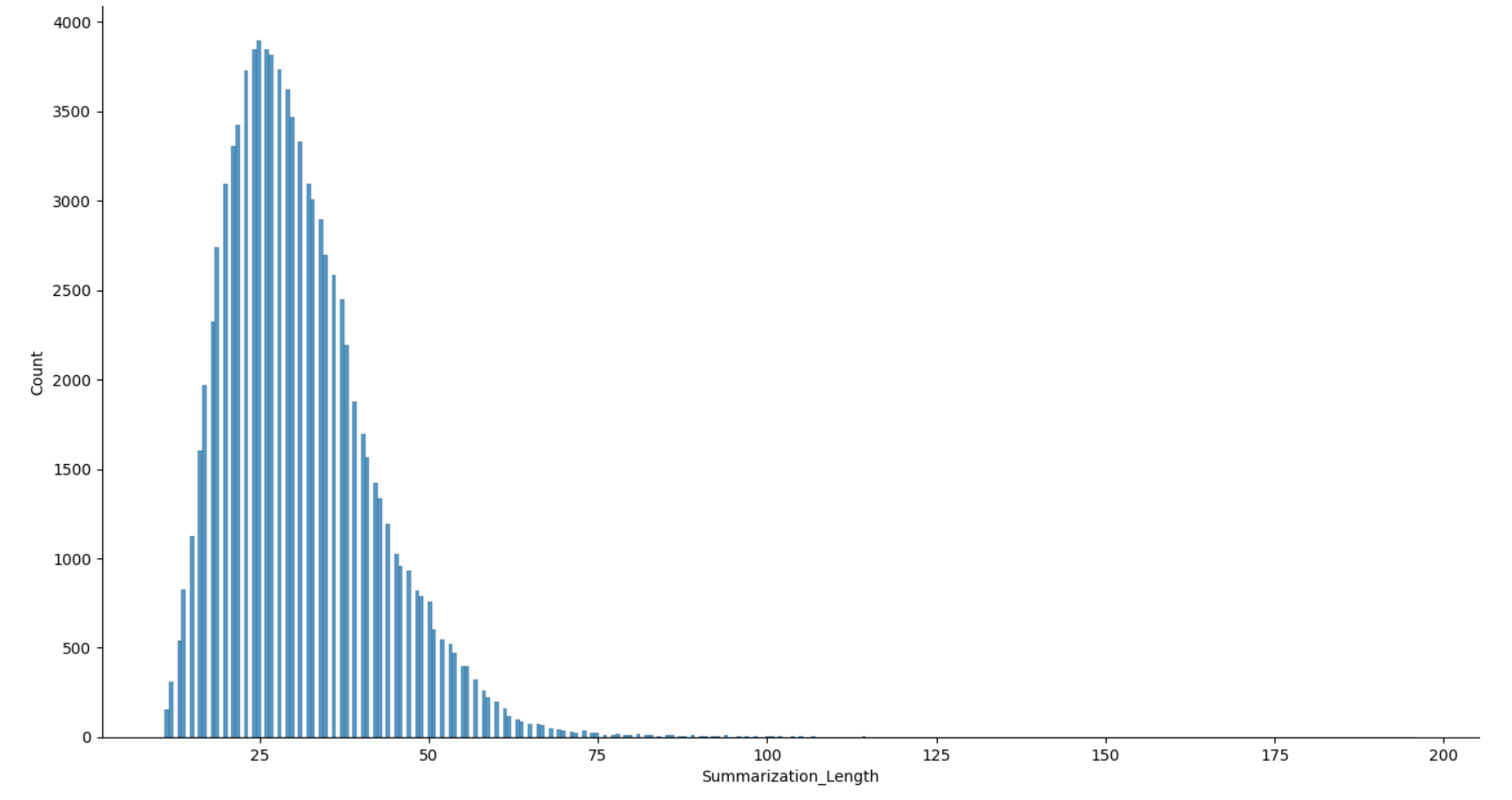
**شکل 3-ابر لغات اخبار موجود در مجموعه داده**

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[[1]](#footnote-1) 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 4 و شکل 5 به ترتیب توزیع تعداد توکن‌های موجود در متن و خلاصه اخبار را نمایش می‌دهند.



**شکل 4-توزیع تعداد کلمات(توکن‌ها) موجود در متن اخبار**



**شکل 5-توزیع تعداد کلمات(توکن‌ها) موجود در خلاصه اخبار**

نمونه‌ای از ستون‌هایی از دادگان که در این پروژه مورد استفاده خواهد بود در شکل 6 آورده شده است.



**شکل 6-چند نمونه از رکوردهای موجود در مجموعه داده نهایی(ستون News ادامه دارد و بخش ابندایی آن نمایش داده شده است).**

1. Hazm Library [↑](#footnote-ref-1)