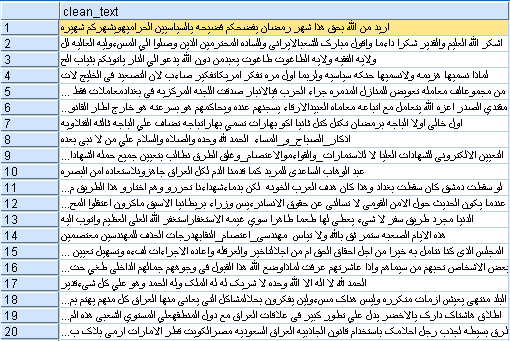
-\*گزارش آماده سازی داده \*-

1. **اتصال به الستیک و استخراج داده**

پاکسازی داده:

* حذف جملات غیر عربی
* حذف کارکترهای اضافی
* حذف حروف غیر از عربی در جملات عربی
* نرمال سازی متن
* انتخاب جملات بیشتر از 75 کاراکتر

تعداد داده جمع شده:

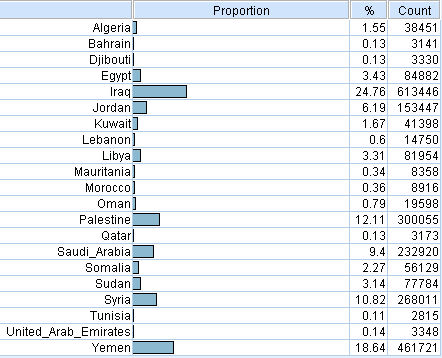
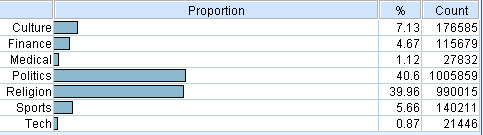
* الستیک سوریه: 1،492،926
* الستیک عراق:1،459،219
* مجموع: 2،952،145

1. **اجرای مدل sentiment, topic, dialect**

داده تحویل گرفته شده 12 فایل مختلف بود که تکراری ها حذف و بعد merge شدند.

تعداد داده ها : 2،477،627

توزیع لیبل ها:

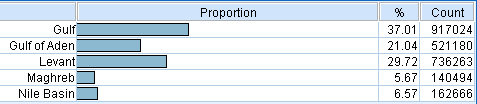


1. **تبدیل 21 کلاس لهجه به 5 کلاس**

با توجه به الگوی زیر کلاس های لهجه به 5 کلاس کاهش یافت

1. مغربی: لیبی، تونس، مراکش، الجزایر و موریتانی
2. شامی: سوریه، اردن، لبنان و فلسطین
3. خلیجی: عراق، قطر، عمان، عربستان، امارات، کویت و بحرین
4. عدنی: یمن، جیبوتی و سومالی
5. نیلی: مصر و سودان

توزیع لهجه 5 کلاسه:

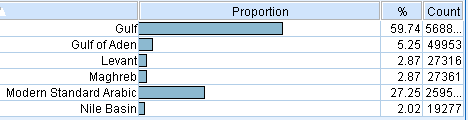


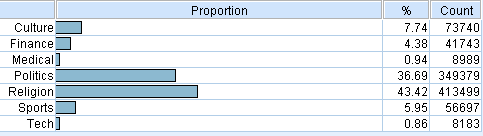
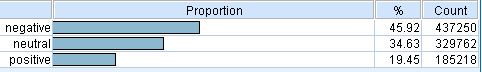
1. **اجرای مدل دوم لهجه (6 کلاسه)**

مدل دوم تشخیص لهجه روی داده ها اعمال شد و داده های با شروط زیر انتخاب شدند:

* داده هایی که هر دو مدل لهجه یکسان پیش بینی شده اند
* داده هایی که مدل دوم به عنوان کلاس لهجه استاندارد انتخاب کرده همان را انتخاب می کنیم

تعداد داده ها: 952،230

توزیع داده ها:



1. **بالانس مجموعه داده**

اقدامات زیر به ترتیب برای بالانس کردن داده ها اننجام شد:

* در لهجه لیبل خلیج را به تعداد لیبل استاندارد **کاهش** دادیم
* در لهجه ما بقی لیبل ها را به تعداد استاندارد **افزایش** دادیم
* در موضوع همه لیبل ها را به تعداد رلیجن **افزایش** دادیم
* در احساس همه لیبل ها را به تعداد لیبل مثبت **کاهش** دادیم
* در لهجه همه لیبل ها را به تعداد لیبل مغرب **کاهش** دادیم

در نهایت 25 آبان 1401

**تعداد داده: 2،176،252**

**توزیع داده:**

