گزارش پروژه: تحلیل و خوشه‌بندی داده‌های شبکه اجتماعی

DataMinds (Zohre Nasiri Zarandi, Moein Zeynodini, Farhan Kian)

۱. مقدمه و هدف پروژه

هدف این پروژه، تحلیل داده‌های شبکه اجتماعی شامل تعداد لایک‌ها، کامنت‌ها، و اشتراک‌گذاری‌ها و همچنین اطلاعات تکمیلی از جمله نرخ تعامل Engagement Rate و داده‌های قیمت دلار، با استفاده از روش‌های **پیش‌پردازش داده، خوشه‌بندی (Clustering) و تحلیل آماری** است.
این تحلیل می‌تواند برای شناسایی الگوهای رفتاری کاربران و ارتباط آن با تغییرات اقتصادی (مانند نوسانات قیمت دلار* مورد استفاده قرار گیرد.

۲. بارگذاری کتابخانه‌ها و داده‌ها

- در ابتدای کد کتابخانه‌های مورد نیاز بارگذاری شده‌اند:
- سپس داده اصلی (Project1_Dataset.csv* خوانده شده و داده‌های منفی از ستون‌های likes, comments, و shares حذف شده‌اند.

۳. پیش‌پردازش داده‌ها

- **حذف مقادیر منفی:** برای اطمینان از معنادار بودن داده‌ها، رکوردهایی که مقدار لایک، کامنت یا اشتراک‌گذاری منفی داشتند حذف شدند.
- **تبدیل تاریخ شمسی به میلادی:** با استفاده از کتابخانه persiantools تاریخ‌ها از فرمت شمسی به میلادی تبدیل شدند تا امکان ادغام با داده‌های قیمت دلار فراهم شود.
- **ادغام داده‌ها:** داده شبکه اجتماعی با داده قیمت دلار بر اساس ستون مشترک date_gregorian ادغام شد.
- **مدیریت فرمت تاریخ شمسی** و ادغام داده‌های اینستاگرام با داده‌های قیمت دلار
- **پر کردن مقادیر گمشده** در ستون‌های مرتبط با دلار و شناسه اینفلوئنسرها
- تبدیل صحیح انواع داده‌ها (به‌ویژه ستون‌های change_amount و change_percent*
- **ایجاد ویژگی‌های جدید** مانند مجموع تعاملات، تعامل به ازای هر دنبال‌کننده و **نوسان قیمت دلار**
- کدگذاری متغیرهای دسته‌ای (مانند category، content_type و influencer_id*
- **مقیاس‌بندی ویژگی‌های عددی** برای استفاده در الگوریتم‌های یادگیری ماشین
- **مدیریت داده‌های پرت** در نرخ تعامل (Engagement Rate*، داده های پرت برای تحلیل بهتر حذف نشدند.

- **ایجاد ویژگی‌های دودویی** برای نرخ تعامل بالا و دنبال‌کنندگان زیاد

۴. مهندسی ویژگی‌ها

- نرخ تعامل: بر اساس لایک‌ها، کامنت‌ها، اشتراک‌گذاری‌ها و تعداد دنبال‌کننده‌ها محاسبه شد.
- کدگذاری ویژگی‌های متنی: با استفاده از Label Encoding به اعداد تبدیل شدند تا در الگوریتم‌ها قابل استفاده باشند.
- توضیح درباره اهمیت کامنت‌ها:
- کامنت‌ها نشان‌دهنده تعامل عمیق‌تر کاربر با محتوا هستند.
- حضور کامنت‌ها موجب افزایش دیده‌شدن پست‌ها در الگوریتم‌های شبکه‌های اجتماعی می‌شود.
- کامنت‌ها ایجاد مکالمه و تقویت جامعه کاربران را به دنبال دارند.
- توضیح درباره اهمیت اشتراک‌گذاری‌ها (Shares*:
- اشتراک‌گذاری‌ها نشانه تایید و ارزش‌گذاری محتوا توسط کاربران است.
- این کار موجب گسترش ارگانیک محتوا در شبکه می‌شود.
- اشتراک‌گذاری می‌تواند اثر ویروسی ایجاد کند و اعتبار اجتماعی محتوا را افزایش دهد.

۵. استانداردسازی داده‌ها

با استفاده از **StandardScaler** ویژگی‌های عددی مقیاس‌بندی شدند تا در فرآیند خوشه‌بندی، همه ویژگی‌ها وزن برابر داشته باشند:

۶. خوشه‌بندی داده‌ها

الگوریتم K-Means با تعداد خوشه ۳ (k=3* برای خوشه‌بندی داده‌ها به کار رفت. هدف از این مرحله، تقسیم داده‌ها به گروه‌هایی با رفتار مشابه در تعاملات و ویژگی‌های دیگر بود.

۷. تحلیل خوشه‌ها

پس از خوشه‌بندی، برای هر خوشه تحلیل آماری و بررسی همبستگی بین نرخ تعامل و تعداد دنبال‌کننده‌ها انجام شد.

Cluster	Features Summary	Suggested Name	Explanation (Persian)
0	- log_follower_count: ~0.98 (medium-high) - engagement_rate: slightly negative (-0.057) - weighted_engagement: slightly negative (-0.065) - likes/comments/shares: slightly negative - is_sponsored: ~0.60 (60% sponsored) - hashtag_count: near zero	**Medium Followers, Moderate Sponsored**	این خوشه شامل اینفلوئنسرهایی است که تعداد فالوئر متوسط تا نسبتاً بالا دارند، نرخ تعامل متوسط به پایین و تقریباً 60 درصد پست‌هایشان اسپانسردار است.
1	- log_follower_count: 0.69 (کمتر از خوشه‌های دیگر) - engagement_rate: خیلی پایین (-0.24*) - weighted_engagement: خیلی پایین (-0.27*) - likes/comments/shares: خیلی پایین - is_sponsored: بالاتر از 60% (0.64*) - hashtag_count: خیلی بالاتر از بقیه (0.53*)	**Low Followers, Low Engagement, High Hashtags & Sponsored**	این خوشه شامل اینفلوئنسرهایی با تعداد فالوئر کم، تعامل بسیار پایین، ولی استفاده زیاد از هشتگ‌ها و درصد زیادی پست اسپانسردار دارند. معمولاً این‌ها کسانی هستند که تلاش می‌کنند با تبلیغات و هشتگ زیاد دیده شوند.
2	- log_follower_count: 0.99 (بالاترین) - engagement_rate: مثبت (0.08*) - weighted_engagement: مثبت (0.09*) - likes/comments/shares: مثبت - is_sponsored: حدود 60% - hashtag_count: نزدیک صفر	**High Followers, High Engagement, Balanced Sponsored**	این گروه اینفلوئنسرهای برتر با بیشترین تعداد فالوئر و نرخ تعامل بالاتر است که پست‌های اسپانسردار متوسط دارند و هشتگ کمتری نسبت به خوشه 1 استفاده می‌کنند.

پس از خوشه‌بندی، برای هر خوشه تحلیل آماری انجام شد و مشخص شد:

خوشه ۲:

- بیشترین تعداد دنبال‌کننده، نرخ تعامل و تعامل وزن‌دار را دارد.
- بیشترین میانگین لایک و کامنت‌ها متعلق به این خوشه است.

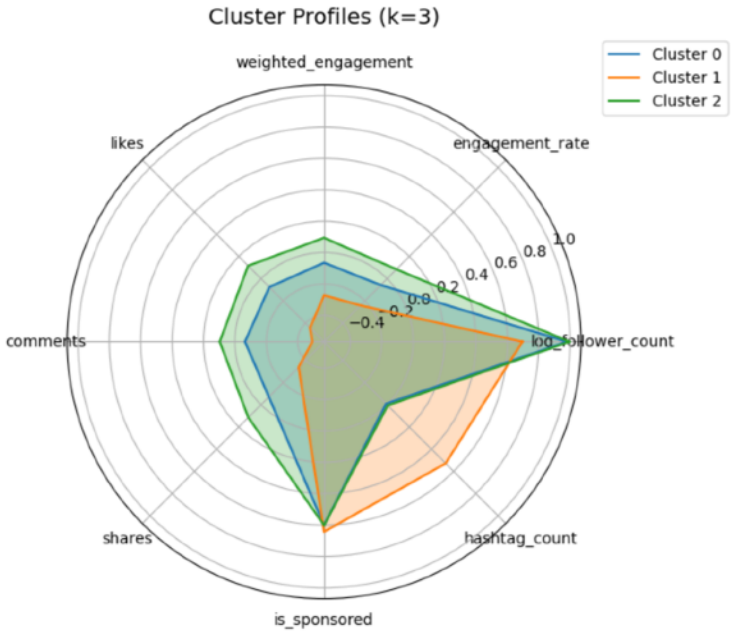
خوشه ۱:

- کمترین تعداد دنبال‌کننده و نرخ تعامل را دارد.
- بیشترین استفاده از هشتگ‌ها در این خوشه مشاهده می‌شود که ممکن است تلاش برای جبران تعامل کمتر باشد.

خوشه ۰:

- در موقعیت متوسط قرار دارد و تعداد دنبال‌کننده و نرخ تعامل آن میانگین دو خوشه دیگر است.
- پارامترهای دیگر مانند اسپانسر بودن (is_sponsored*) تقریبا در هر سه خوشه مشابه است.

نمودار زیر یک **رادار چارت (Radar Chart)** از پروفایل سه خوشه (Cluster 0*، Cluster 1، Cluster 2) را نشان می‌دهد که حاصل **خوشه‌بندی k=3** است.



تحلیل بر اساس ویژگی‌ها:

1. **log_follower_count**: خوشه 2 ، سبز بیشترین تعداد دنبال‌کننده را دارد، کمی بالاتر از خوشه 0 ، آبی و خوشه 1 ، نارنجی دنبال‌کننده‌های کمتری نسبت به بقیه دارد.
2. **engagement_rate** و **weighted_engagement**: خوشه 2 به وضوح نرخ تعامل بالاتری دارد و خوشه 1 کمترین نرخ تعامل را دارد.
3. **likes و comments**: خوشه 2 بیشترین لایک و کامنت را دریافت می‌کند و خوشه 1 بسیار پایین است، مخصوصاً در کامنت‌ها که تقریباً صفر است.
4. **shares**: خوشه 2 و 0 بالاتر از خوشه 1 هستند، ولی تفاوت شدید نیست.
5. **is_sponsored**: هر سه خوشه تقریباً یکسان هستند، احتمالاً به این معنی که اسپانسر بودن در همه گروه‌ها به نسبت مشابه رخ می‌دهد.
6. **hashtag_count**: خوشه 1 بیشترین استفاده از هشتگ را دارد و خوشه 0 کمترین تعداد هشتگ را استفاده می‌کند.

نتیجه گیری :

- خوشه ۲ بهترین عملکرد را دارد: بیشترین دنبال‌کننده، بیشترین نرخ تعامل و بیشترین تعامل وزن‌دار. این یعنی این دسته از اینفلوئنسرها، فالوورهای زیادی دارند و تعامل بیشتری هم با پست‌هایشان اتفاق می‌افتد.
- خوشه ۱ تعداد دنبال‌کننده کمتر و نرخ تعامل پایین‌تری دارد ولی بیشترین تعداد هشتگ را استفاده می‌کند. ممکن است این گروه برای جبران تعامل کمتر، از هشتگ‌های بیشتری استفاده کنند.
- خوشه ۰ در میان دو خوشه دیگر است؛ تعداد دنبال‌کننده و تعامل متوسط دارد.

۸. نتایج آماری

همبستگی بین تعداد دنبال‌کننده و نرخ تعامل: محاسبات نشان داد که تقریباً هیچ رابطه‌ی خطی بین تعداد فالوورها و نرخ تعامل وجود ندارد. مقدار همبستگی بسیار نزدیک به صفر ، ۰.۰۰۲۸ بود و مقدار p-value برابر ۰.۹۱۳۷ که نشان‌دهنده عدم وجود رابطه خطی معنادار بین این دو متغیر است. این یعنی تعداد دنبال‌کننده به تنهایی تعیین‌کننده نرخ تعامل نیست و عوامل دیگری مانند نوع محتوا یا زمان انتشار نقش مهم‌تری دارند.

این موضوع ممکن است نشان دهد که میزان تعامل لزوماً به تعداد فالوور وابسته نیست و عوامل دیگری مثل نوع محتوا، تعامل واقعی کاربران، یا زمان انتشار پست مهم‌تر هستند.

آزمون ANOVA برای مقایسه نرخ تعامل بین انواع محتوا، تصویر، ویدیو، چرخشی، متن: آزمون ANOVA تحلیل واریانس بررسی می‌کند که آیا میانگین متغیر عددی ، در اینجا engagement rate بین چند گروه مختلف ، در اینجا انواع محتوای text ، carousel ، video ، image تفاوت معنادار آماری دارد یا نه.

در داده‌ها مشاهده شد، نوع محتوای منتشرشده ، عکس، ویدیو، چرخشی یا متنی تأثیر مشخصی بر نرخ تعامل نداشته است. به بیان دیگر، نرخ تعامل بین این دسته‌ها تفاوت آماری قابل‌توجهی ندارد و ممکن است این تفاوت‌ها تصادفی باشند. نتایج نشان داد که تفاوت معنادار آماری بین میانگین نرخ تعامل این گروه‌ها وجود ندارد. یعنی نوع محتوای منتشر شده تاثیر مشخص و قابل توجهی بر نرخ تعامل نداشته است.

۹. جمع‌بندی

در این پروژه:

- داده‌های شبکه اجتماعی پاکسازی و به داده‌های قیمت دلار متصل شدند.
- ویژگی‌های کلیدی استخراج و مقیاس‌بندی شدند.
- الگوریتم خوشه‌بندی K-Means به خوبی داده‌ها را به سه گروه مجزا تقسیم کرد.
- تحلیل‌های آماری نشان داد که تعداد دنبال‌کننده‌ها به تنهایی معیار مناسبی برای پیش‌بینی نرخ تعامل نیست.
- نوع محتوا نیز تفاوت معناداری در نرخ تعامل ایجاد نکرده است.
- یافته‌ها می‌توانند در بهبود استراتژی‌های بازاریابی و مدیریت شبکه‌های اجتماعی کاربردی باشند.