بسم الله الرحمن الرحیم

آنالیز علامت ها و ایموجی برای محصولات دیجیکالا

تیم ارائه دهنده: داده کاوانِ فونیکس

ناظر و سرپرست تیم: دکتر علیرضا وفایی صدر

درخواست دهنده: ستاد علوم شناختی IPM

زمستان ۱۳۹۸

.

تعریف پروژه

در این گزارش، به بررسی تعداد علامت ها و سمبل های شناختی و تعداد ایموجی ها در داده های دیجیکالا پرداختیم. این کار توسط *آرمیتا رضوی* بر روی داده های اصلی و پیش پردازش نشده و توسط *خانم ستاره* ترنج بر روی داده های پاکسازی شده انجام گرفت.

بررسی ایموجی و سمبل ها

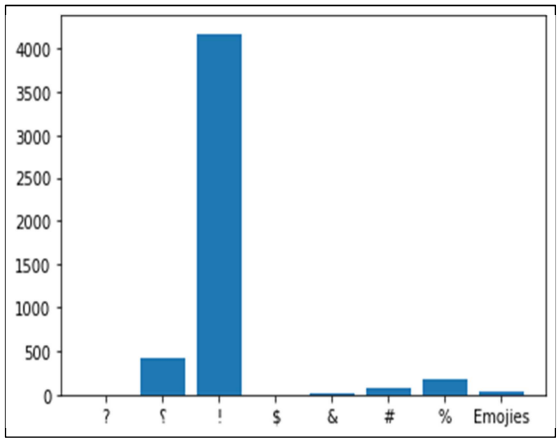
در ستون های کامنت، عنوان، مزایا و معایب محصول، شروع به پشتن علامت های مختلف و ایموجی کردیم و نتایج به صورت زیر حاصل شد:

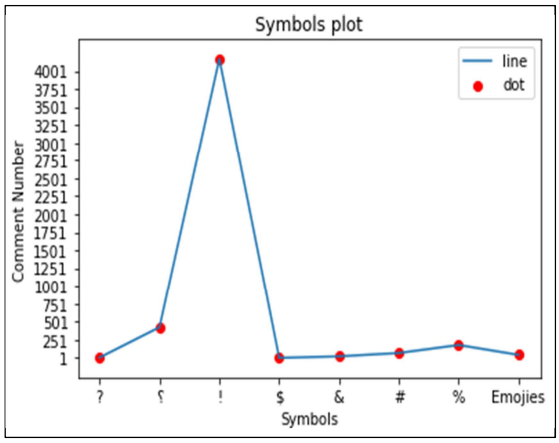
برای داده های اصلی و پردازش نشده دیجیکالا با دیتاست Digi_2.xlsx :

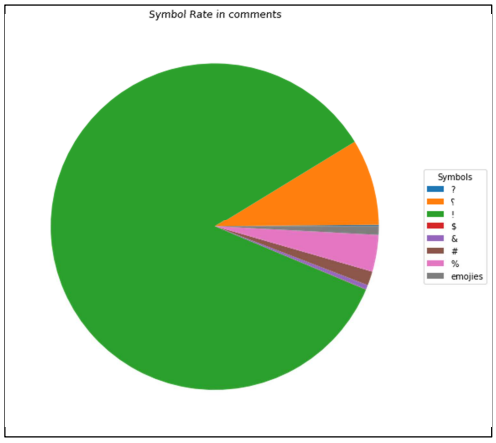
علامت های مورد بررسی:

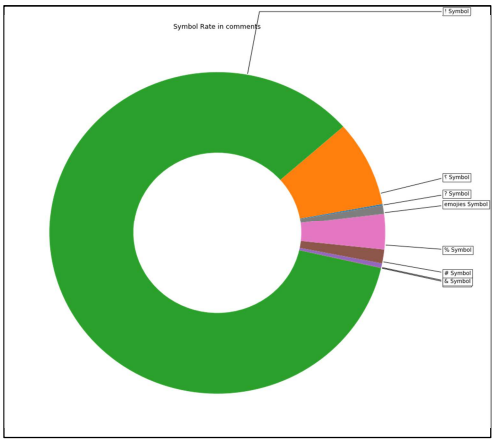
؟ ! % # & $? و ایموجی ها

نمودارهای زیر نشان دهنده تعداد کامنت های موجود که شامل علائم مختلف می باشند است:

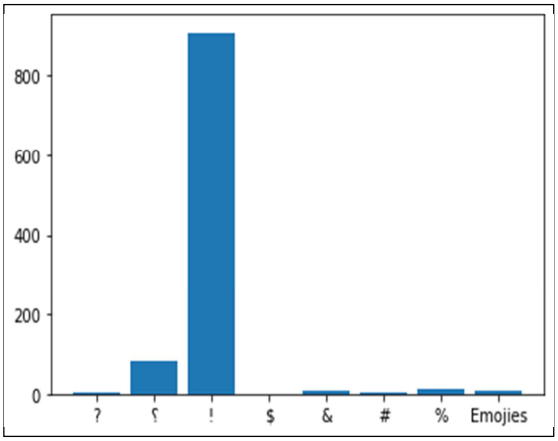


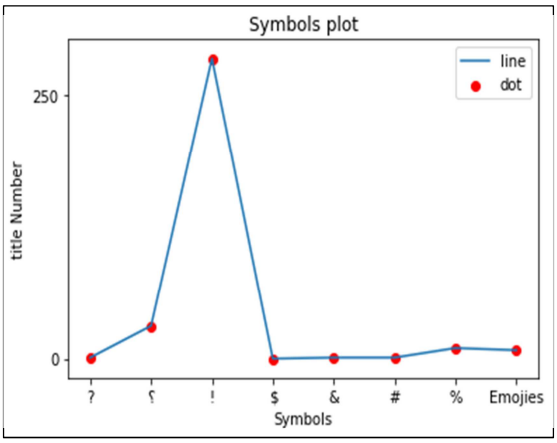


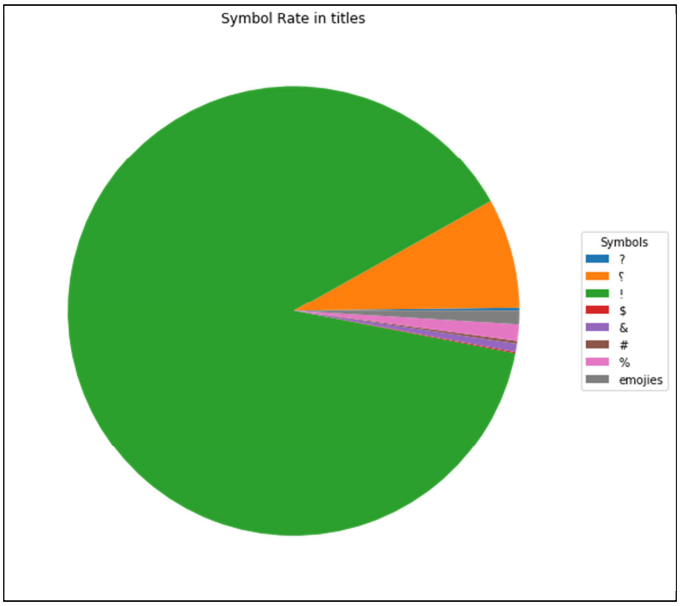




نمودارهای زیر نشان دهنده تعداد عنوان کامنت موجود که شامل علائم مختلف می باشند است:







همینطور برای ستون های مزایا و معایب نیز این پلات ها کشیده شده است. جدول زیر تعداد داده های در ستون مختلف که شامل علامت ها و ایموجی هستند را نشان می دهد:

Symbol	Comment count	Title count	Advantages count	Disadvantages count
؟	6	2	1	1
؟	423	81	11	31
!	4169	908	143	284
$	1	1	0	0
&	22	6	5	1
#	68	2	0	1
%	178	12	12	10
Emoji	42	10	2	8

Number of Symbols in Dataset Columns

فایل این آنالیز در مسیر زیر قرار دارد:

Report2->Armita Razavi -> Find_Emojies-Symbols-ipynb

همچنین تمامی عکس ها وپلات ها در فولدری در مسیر زیر موجود است:

Report2->Armita Razavi -> Symbols_Not_Clean

همانطور که مشاهده می شود بیشتر سطرها در ستون های مختلف علامت تعجب و کترین علامت مربوط به علامت $ می باشد. با توجه به آمار، کمتر از ۱۰٪ از داده های و سطرها شامل علایم شناختی و ایموجی ها هستند که اصلا به چشم نمیاد.

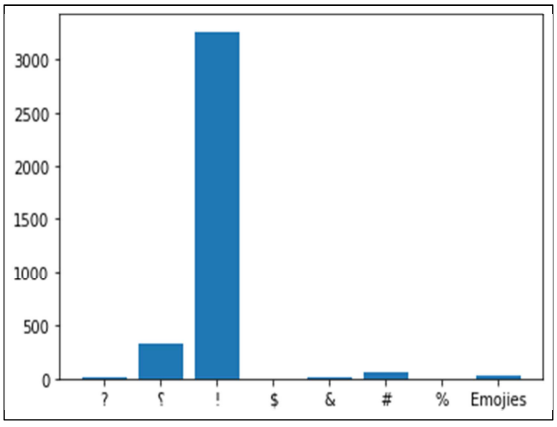
در ستون های کامنت، عنوان، مزایا و معایب محصول، شروع به پشتن علامت های مختلف و ایموجی کردیم و نتایج به صورت زیر حاصل شد:

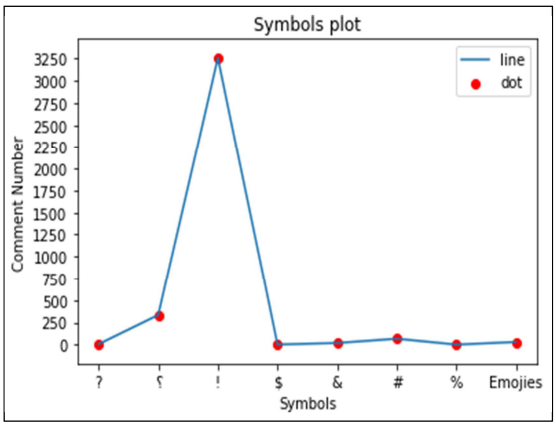
برای داده های پاکسازی شده ولی با حفظ علائم و ایموجی ها دیجیکالا با دیتاست digi_clean_final.xlsx :

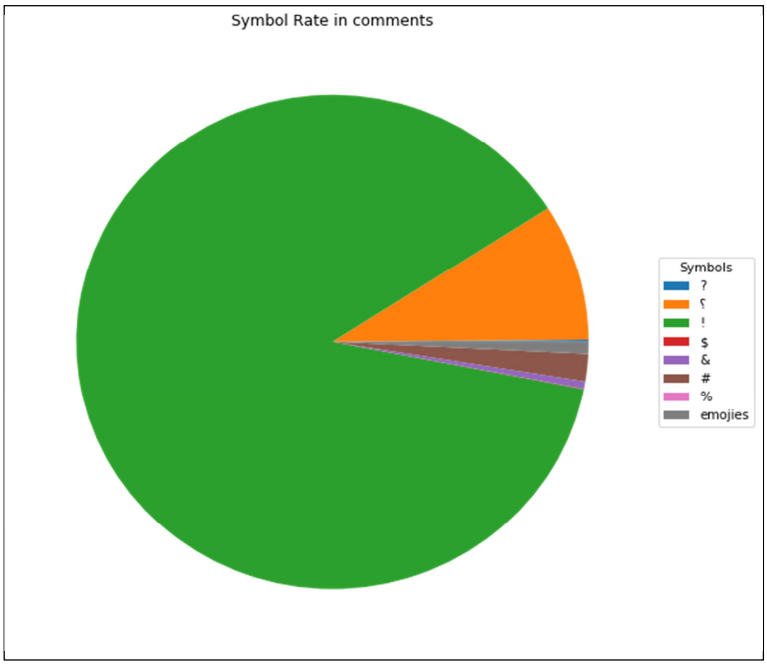
علامت های مورد بررسی:

؟ ! % # & $? و ایموجی ها

نمودارهای زیر نشان دهنده تعداد کامنت های موجود که شامل علائم مختلف می باشند است:







. جدول زیر تعداد داده های در ستون مختلف که شامل علامت ها و ایموجی هستند را بعد از پردازش داده ها نشان می دهد:

Symbols	Comment	Title	Advantages	Disadvantages
?	5	2	1	0
؟	332	55	10	26
!	3259	693	112	225
$	1	1	0	0
&	18	4	4	1
#	66	376	1104	2860
%	0	0	0	0
emoji	29	5	2	7

فایل و پلات های این بخش در مسیر Report2->Setare Toranj -> Symbol قرار گرفته است.

کلیه فایل های مربوط به این گزارش در گیت هاب زیر و در فولدر **Report2** موجود است.

https://github.com/phoenix-dataminers/Digikala2

پایان