کارگاه اول : یادگیری ماشین

اعضای گروه : فاطمه قربانی ،زهرا خداقلی ،بهاره فرهنگ

آنالیز

این مجموعه داده شامل ویژگی های زیر است.

0. url: آدرس خبر
1. timedelta: تعداد روزها بین پخش خبر و تولید دیتاست
2. n_tokens_title: تعداد کلمات در عنوان
3. n_tokens_content: تعداد کلمات در متن
4. n_unique_tokens: نرخ کلمات خاص در متن
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: تعداد لینک ها
8. num_self_hrefs: تعداد لینک ها به مقاله های دیگر درهمین سایت
9. num_imgs: تعداد تصاویر
10. num_videos: تعداد ویدیو ها
11. average_token_length: میانگین طول کلمات
12. num_keywords: Number of keywords in the metadata
13. data_channel_is_lifestyle: موضوع سبک زندگی
14. data_channel_is_entertainment: موضوع تفریحات
15. data_channel_is_bus: موضوع تجارت
16. data_channel_is_socmed: موضوع شبکه اجتماعی
17. data_channel_is_tech: موضوع تکنولوژی
18. data_channel_is_world: موضوع جهان
19. kw_min_min: Worst keyword (min. shares)
20. kw_max_min: Worst keyword (max. shares)
21. kw_avg_min: Worst keyword (avg. shares)
22. kw_min_max: Best keyword (min. shares)
23. kw_max_max: Best keyword (max. shares)
24. kw_avg_max: Best keyword (avg. shares)
25. kw_min_avg: Avg. keyword (min. shares)
26. kw_max_avg: Avg. keyword (max. shares)
27. kw_avg_avg: میانگین کلمات کلیدی(میانگین اشتراک گذاری)
28. self_reference_min_shares: Min. shares of referenced articles in Mashable
29. self_reference_max_shares: Max. shares of referenced articles in Mashable
30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
31. weekday_is_monday: آیا دوشنبه منتشر شده؟
32. weekday_is_tuesday: آیا سه شنبه منتشر شده؟
33. weekday_is_wednesday: آیا چهار شنبه منتشر شده؟
34. weekday_is_thursday: آیا پنج شنبه منتشر شده؟
35. weekday_is_friday: آیا جمعه منتشر شده؟
36. weekday_is_saturday: آیا شنبه منتشر شده؟
37. weekday_is_sunday: آیا یکشنبه منتشر شده؟
38. is_weekend: آیا آخر هفته منتشر شده؟
39. LDA_00: Closeness to LDA topic 0
40. LDA_01: Closeness to LDA topic 1
41. LDA_02: Closeness to LDA topic 2
42. LDA_03: Closeness to LDA topic 3
43. LDA_04: Closeness to LDA topic 4
44. global_subjectivity: Text subjectivity
45. global_sentiment_polarity: Text sentiment polarity

46. global_rate_positive_words: نرخ کلمات مثبت در متن
47. global_rate_negative_words: نرخ کلمات منفی در متن
48. rate_positive_words: Rate of positive words among non-neutral tokens
49. rate_negative_words: Rate of negative words among non-neutral tokens
50. avg_positive_polarity: Avg. polarity of positive words
51. min_positive_polarity: Min. polarity of positive words
52. max_positive_polarity: Max. polarity of positive words
53. avg_negative_polarity: Avg. polarity of negative words
54. min_negative_polarity: Min. polarity of negative words
55. max_negative_polarity: Max. polarity of negative words
56. title_subjectivity: Title subjectivity
57. title_sentiment_polarity: Title polarity
58. abs_title_subjectivity: Absolute subjectivity level
59. abs_title_sentiment_polarity: Absolute polarity level
60. shares: تعداد اشتراک گذاری خبر

میانگین اشتراک گذاری ها 3395.38 با انحراف معیار 11626.95 و میانه ۱۴۰۰ است. و با توجه به این که کمترین آنها ۱ و بیشترین 843300 است پس میانگین معیار درستی برای گذاشتن threshold نیست و میزان میانه را برای اینکار درنظر میگیریم.

Popular = shares >1400

Unpopular = shares <= 1400

با توجه به اینکه ویژگی آدرس خبر و تعداد روز ها تاثیری در اشتراک گذاری خبر که هدف مدل است ندارد پس بهتر است این ویژگی ها را حذف کنیم.

ابتدا نموداری برای مقایسه اهمیت روزهای هفته میکشیم و چون این ویژگی در ۷ ویژگی جدا پخش شده باید آنها را یکی کنیم. به نظر میرسد در اواسط هفته احتمال آنکه خبر زیاد به اشتراک گذاشته شود کمتر است و این میزان در روزهای پایانی هفته به بیشترین میزان خود میرسد. هر چند مجموع کل اخبار در این روزها کمتر است.

نمودار دوم برای مقایسه اهمیت موضوع خبر است. موضوعات جهان و تفریحات از اشتراک گذاری کمتری برخوردارند و اخبار مربوط به شبکه های اجتماعی و تکنولوژی از بیشترین اشتراک گذاری برخوردارند.

با مشاهده نمودار همبستگی میبینیم که بیشترین همبستگی بین ۱۵ ویژگی زیر است.

- kw_avg_avg
- LDA_03
- kw_max_avg
- kw_min_avg
- num_hrefs
- num_imgs
- self_reference_avg_sharess
- is_weekend
- self_reference_min_shares
- self_reference_max_shares
- kw_avg_max
- global_subjectivity

- abs_title_sentiment_polarity
- weekday_is_sunday
- title_subjectivity

که نمودار آنها را رسم میکنیم.و در هنگام آموزش مدل این ویژگی ها را به تنهایی امتحان میکنیم.

Naïve bayes

اجرای این الگوریتم روی تمام ویژگی ها عملکرد زیر را داشت.

```
Confusion Matrix:
[[5036  549]
 [5062 1247]]
Classification Report:
              precision    recall  f1-score   support

           0       0.50      0.90      0.64      5585
           1       0.69      0.20      0.31      6309

    accuracy                           0.53     11894
   macro avg       0.60      0.55      0.47     11894
weighted avg       0.60      0.53      0.46     11894

Accuracy: 0.5282495375819741
```

اما فقط با اجرا روی بعضی ویژگی ها مثل روزهای هفته و موضوع خبر عملکرد اندکی بهتر شد.

```
Confusion Matrix:
[[4229 1356]
 [3586 2723]]
Classification Report:
              precision    recall  f1-score   support

           0       0.54      0.76      0.63      5585
           1       0.67      0.43      0.52      6309

    accuracy                           0.58     11894
   macro avg       0.60      0.59      0.58     11894
weighted avg       0.61      0.58      0.57     11894

Accuracy: 0.5844963847317975
```

اما بهترین عملکرد استفاده از ویژگی هایی بود که همبستگی بیشتری داشتند.

```
Confusion Matrix:
[[3957 1628]
 [2901 3408]]
Classification Report:
              precision    recall  f1-score   support

           0       0.58      0.71      0.64      5585
           1       0.68      0.54      0.60      6309

    accuracy                           0.62     11894
   macro avg       0.63      0.62      0.62     11894
weighted avg       0.63      0.62      0.62     11894

Accuracy: 0.6192197746763074
```

SVM

اجرای این الگوریتم روی تمام ویژگی به شکل زیر عمل میکند.

```
Confusion Matrix:
[[5171  414]
 [5694  615]]
Classification Report:
              precision    recall  f1-score   support

           0       0.48      0.93      0.63      5585
           1       0.60      0.10      0.17      6309

    accuracy                           0.49     11894
   macro avg       0.54      0.51      0.40     11894
weighted avg       0.54      0.49      0.38     11894

Accuracy: 0.4864637632419707
```

اما اگر فقط ویژگی های روز هفته و موضوع را در نظر بگریم دقت کاهش پیدا میکند.

```
Confusion Matrix:
[[5430  155]
 [6040  269]]
Classification Report:
              precision    recall  f1-score   support

           0       0.47      0.97      0.64      5585
           1       0.63      0.04      0.08      6309

    accuracy                           0.48     11894
   macro avg       0.55      0.51      0.36     11894
weighted avg       0.56      0.48      0.34     11894

Accuracy: 0.47914915083235243
```

با اضافه کردن بعضی ویژگی ها مثل تعداد کلمات دقت اندکی کاهش پیدا میکند اما با زیر مجموعه ای از ویژگی ها با بیشترین همبستگی دقت تقریبا ۵ درصد افزایش میابد.

```
Confusion Matrix:
[[  48 5537]
 [  61 6248]]
Classification Report:
              precision    recall  f1-score   support

           0       0.44      0.01      0.02      5585
           1       0.53      0.99      0.69      6309

    accuracy                           0.53     11894
   macro avg       0.49      0.50      0.35     11894
weighted avg       0.49      0.53      0.37     11894

Accuracy: 0.5293425256431814
```