SHAP analysis

In [ ]:
```
import joblib
from pathlib import Path
from omegaconf import OmegaConf

import pandas as pd
import shap
from sklearn.metrics import classification_report
```

In [ ]:
```
EXPERIMENT_ROOT = "../../experiments/rf_features_only"
RESULTS = Path(EXPERIMENT_ROOT) / "results.yaml"

results = OmegaConf.load(RESULTS)
```

In [ ]:
```
HAS_SCALER = False
DATA_ROOT = "../../data/prepared/"
MODEL = Path(EXPERIMENT_ROOT) / "model.pkl"
SCALER = Path(EXPERIMENT_ROOT) / "scaler.pkl"
VAL_DATA = Path(DATA_ROOT) / "val_features.pkl"
TEST_DATA = Path(DATA_ROOT) / "test_features.pkl"
VEC_COLS = list(range(768))

val_df = pd.read_pickle(VAL_DATA)
val_df_labels = val_df.retweet_label
val_df.drop(["retweet_label", "id_str"], axis=1, inplace=True)

test_df = pd.read_pickle(TEST_DATA)
test_df_labels = test_df.retweet_label
test_df.drop(["retweet_label", "id_str"], axis=1, inplace=True)

model = joblib.load(MODEL)

if HAS_SCALER:
    scaler = joblib.load(SCALER)
    transformed_val = scaler.transform(val_df[VEC_COLS].values)
    val_df[VEC_COLS] = transformed_val
    transformed_test = scaler.transform(test_df[VEC_COLS].values)
    test_df[VEC_COLS] = transformed_test
```

In [ ]:
```
val_df.head()
```

Out[ ]:

	entities.urls	entities.media	user_in_net	has_covid_keyword	tweets_keywords_3_in_deg
173185	0	1	0	0	-0.647
173186	0	1	0	0	-0.647
173187	0	0	0	0	-0.647
173188	0	1	0	0	-0.647
173189	0	1	0	0	-0.647

5 rows × 49 columns

In [ ]:
```
test_df.head()
```

Out[ ]:

	entities.urls	entities.media	user_in_net	has_covid_keyword	tweets_keywords_3_in_deg
194715	1	1	1	1	1.908
194716	1	1	1	0	-0.106
194717	1	1	1	0	-0.310
194718	1	1	1	1	-0.206
194719	1	0	1	0	-0.647

5 rows × 49 columns

Validation and test results

In [ ]:
```
val_predictions = model.predict(val_df.values)
val_out = classification_report(val_df_labels.values, val_predictions,
                                digits=3, output_dict=False)
print(val_out)
```

```
              precision    recall  f1-score   support

           0      0.726     0.674     0.699     10954
           1      0.634     0.690     0.661      8989

    accuracy                          0.681     19943
   macro avg      0.680     0.682     0.680     19943
weighted avg      0.685     0.681     0.682     19943

[Parallel(n_jobs=24)]: Using backend ThreadingBackend with 24 concurrent wo
rkers.
[Parallel(n_jobs=24)]: Done   2 tasks      | elapsed:    0.0s
[Parallel(n_jobs=24)]: Done 152 tasks      | elapsed:    0.1s
[Parallel(n_jobs=24)]: Done 200 out of 200 | elapsed:    0.1s finished
```

In [ ]:
```
test_predictions = model.predict(test_df.values)
test_out = classification_report(test_df_labels.values, test_predictions,
                                 digits=3, output_dict=False)
print(test_out)
```

```
              precision    recall  f1-score   support

           0      0.716     0.640     0.676     10639
           1      0.633     0.710     0.669      9305

    accuracy                          0.673     19944
   macro avg      0.675     0.675     0.673     19944
weighted avg      0.677     0.673     0.673     19944

[Parallel(n_jobs=24)]: Using backend ThreadingBackend with 24 concurrent wo
rkers.
[Parallel(n_jobs=24)]: Done   2 tasks      | elapsed:    0.0s
[Parallel(n_jobs=24)]: Done 152 tasks      | elapsed:    0.1s
[Parallel(n_jobs=24)]: Done 200 out of 200 | elapsed:    0.1s finished
```

SHAP Explainer preparation and test data sampling

In [ ]:
```
explainer = shap.Explainer(model)
```

In [ ]:
```
len(test_df)
```

Out[ ]:
```
19944
```

In [ ]:
```
# sample for faster SHAP calculation
# typically, 100-1000 examples is ok
test_df_sample = test_df.sample(frac=0.05, random_state=42)
```

In [ ]:
```
test_df_labels_sample = test_df_labels[test_df_sample.index]
```

In [ ]:
```
test_df_sample
```

Out[ ]:

	entities.urls	entities.media	user_in_net	has_covid_keyword	tweets_keywords_3_in_deg
196995	0	1	1	0	-0.647
208764	0	1	0	0	0.639
198537	1	0	1	1	0.171
199457	0	0	1	0	-0.647
204573	0	0	1	0	2.409
...	...	...	...	...	
213637	0	1	1	0	1.427
214221	1	0	1	0	-0.206
212261	0	0	1	0	0.257
213929	0	0	1	0	-0.418
198449	1	0	0	0	-0.418

997 rows × 49 columns

In [ ]:
```
test_df_labels_sample
```

Out[ ]:
```
196995    1
208764    1
198537    0
199457    0
204573    0
         ..
213637    1
214221    1
212261    1
213929    1
198449    1
Name: retweet_label, Length: 997, dtype: int64
```

In [ ]:
```
test_df_conf_sample = model.predict_proba(test_df_sample.values)
test_df_conf_sample
```

```
[Parallel(n_jobs=24)]: Using backend ThreadingBackend with 24 concurrent wo
rkers.
[Parallel(n_jobs=24)]: Done   2 tasks      | elapsed:    0.0s
[Parallel(n_jobs=24)]: Done 152 tasks      | elapsed:    0.0s
[Parallel(n_jobs=24)]: Done 200 out of 200 | elapsed:    0.1s finished
```

Out[ ]:
```
array([[0.5306387 , 0.4693613 ],
       [0.39770972, 0.60229028],
       [0.77102733, 0.22897267],
       ...,
       [0.40260657, 0.59739343],
       [0.58668253, 0.41331747],
       [0.70577764, 0.29422236]])
```

In [ ]:
```
shap_values = explainer(test_df_sample)
```

In [ ]:
```
shap_values.base_values.shape
```

Out[ ]:
```
(997, 2)
```

In [ ]:
```
shap_values.values.shape
```

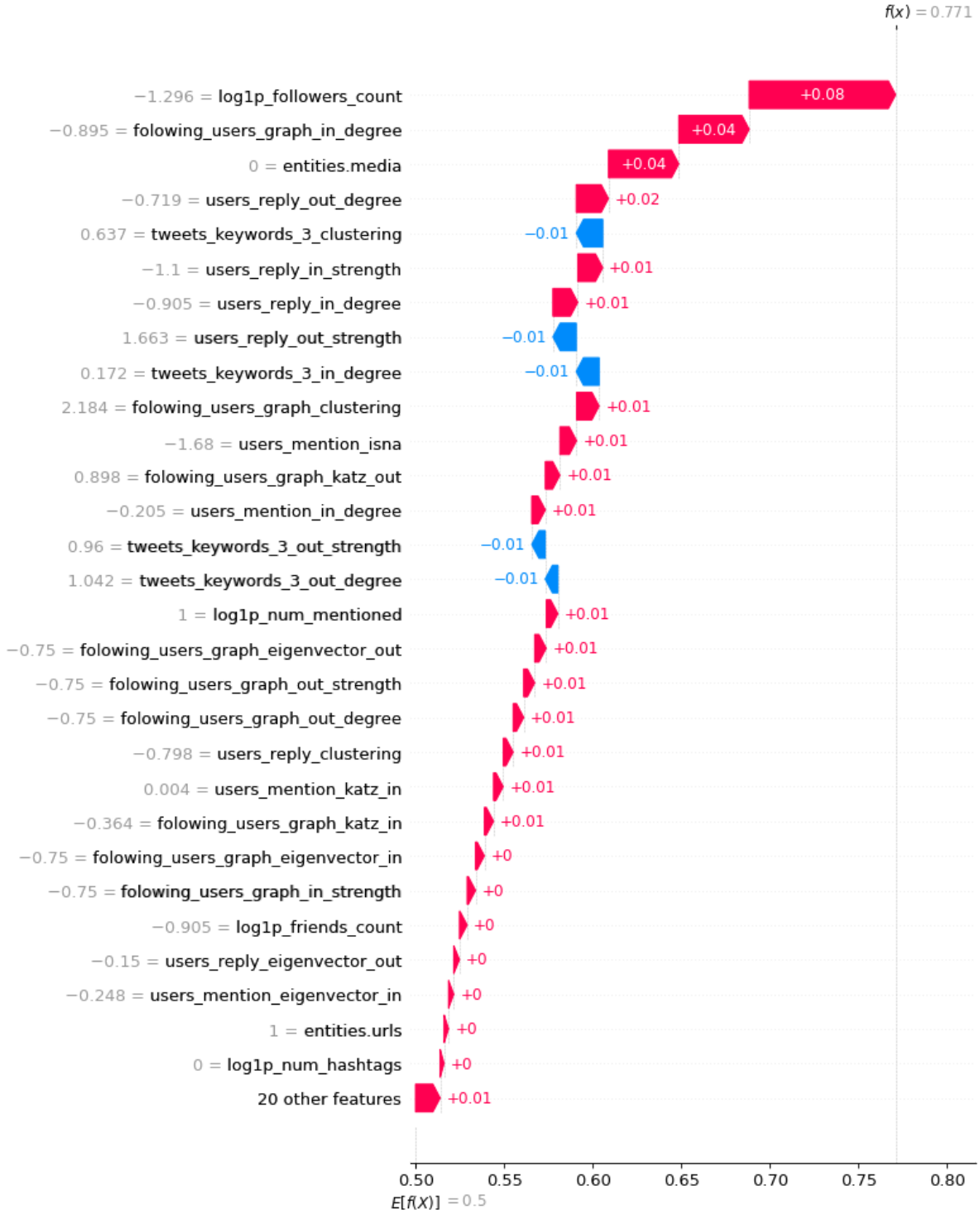
Out[ ]:
```
(997, 49, 2)
```

In [ ]:
```
shap_values.data.shape
```

Out[ ]:
```
(997, 49)
```

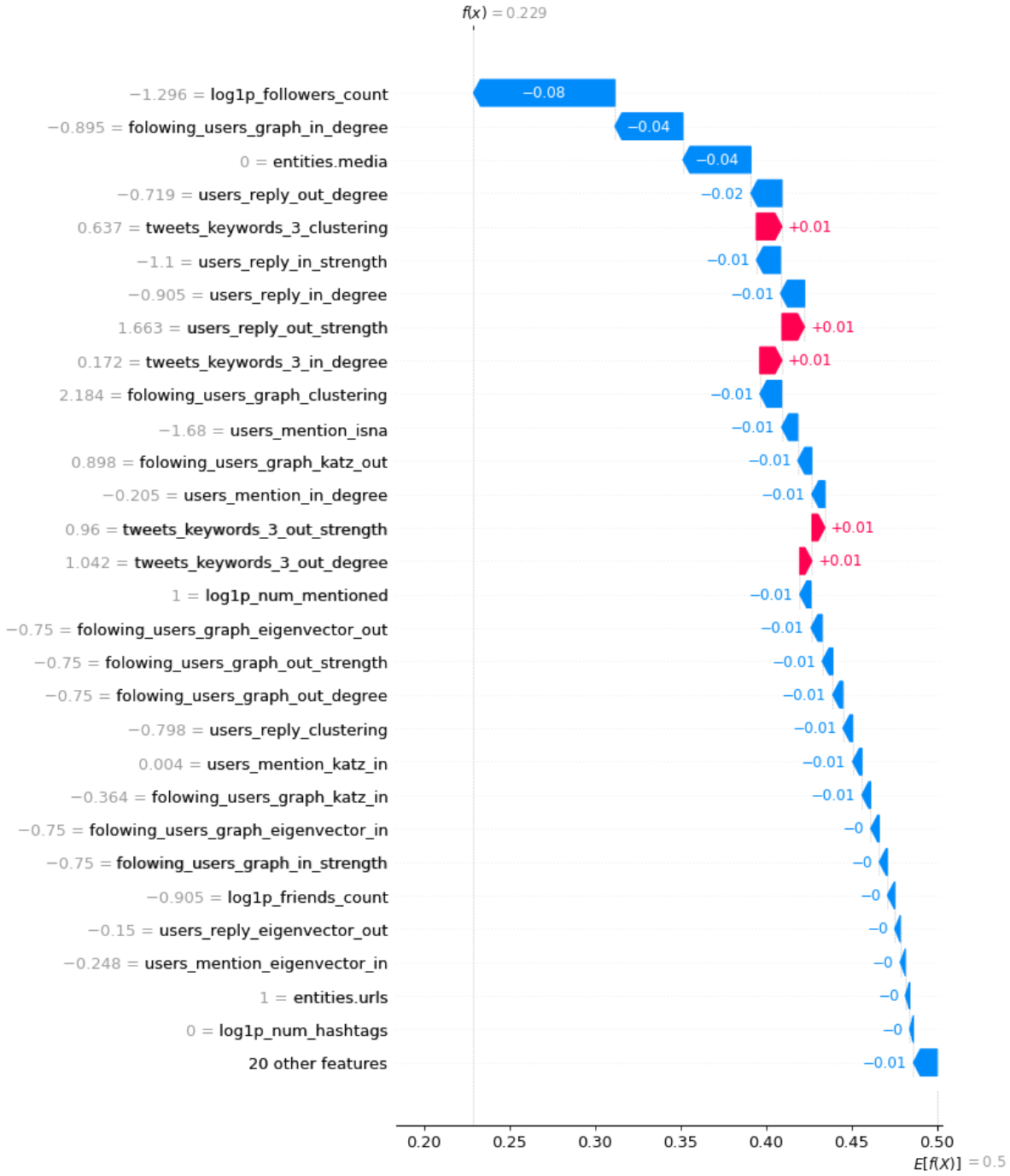
SHAP values accross the test data

In [ ]:
```
# visualize the prediction's explanation for class 0 for a confident correc
# note idx 2 in test_df_conf_sample: 0.771 confidence
idx = 2
exp = shap.Explanation(shap_values.values[:, :, 0], shap_values.base_values
shap.plots.waterfall(exp[idx], max_display=30)
```



In [ ]:

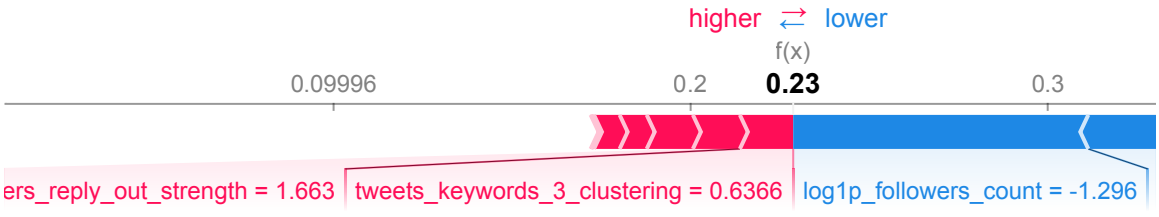
```
# visualize the prediction's explanation for class 1 as a check (symmetric
idx = 2  # the same example
exp = shap.Explanation(shap_values.values[:, :, 1], shap_values.base_values
shap.plots.waterfall(exp[idx], max_display=30)
```



In [ ]:

```
# the same plot for class 1, but horizontal display
idx = 2
shap.initjs()
shap.force_plot(explainer.expected_value[1], shap_values.values[idx, :, 1],
```

Out[ ]:

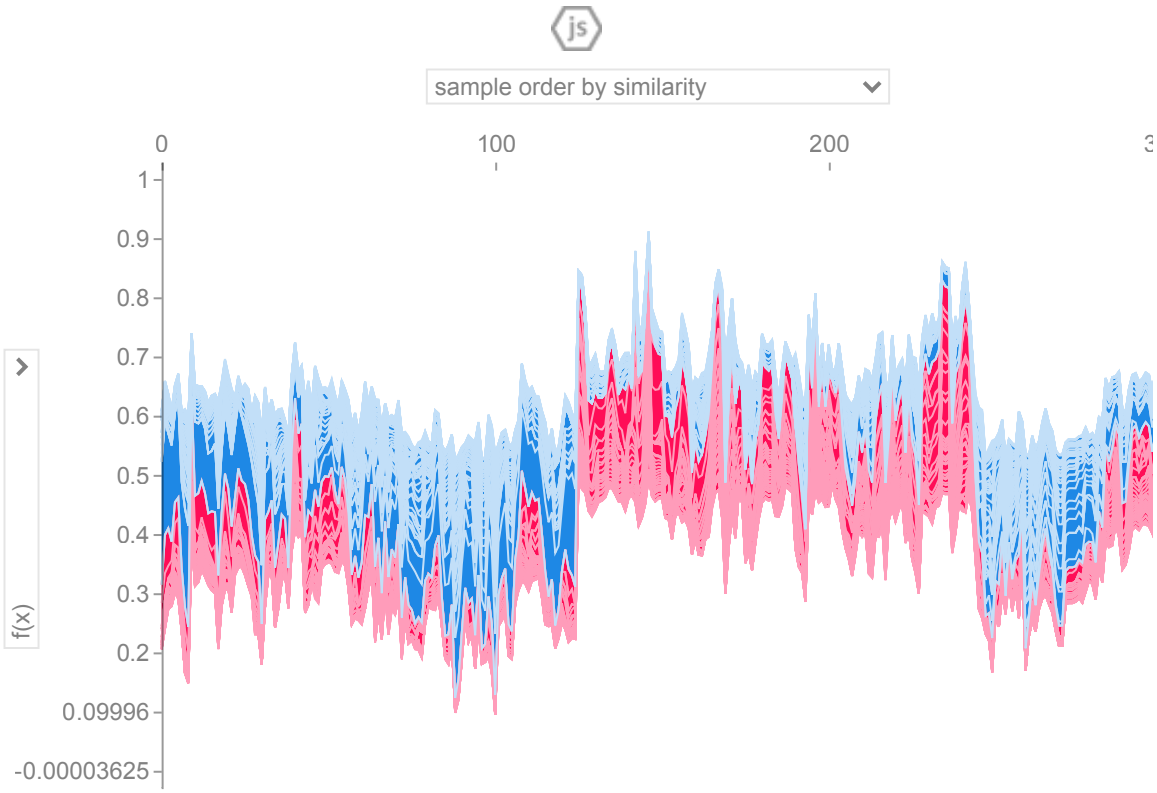


In [ ]:

```
# Vizualize multiple predictions (class 1)
# It is possible to explore different variables interactively in the notebo
```

```
shap.initjs()
shap.force_plot(explainer.expected_value[1], shap_values.values[:, :, 1], t
```

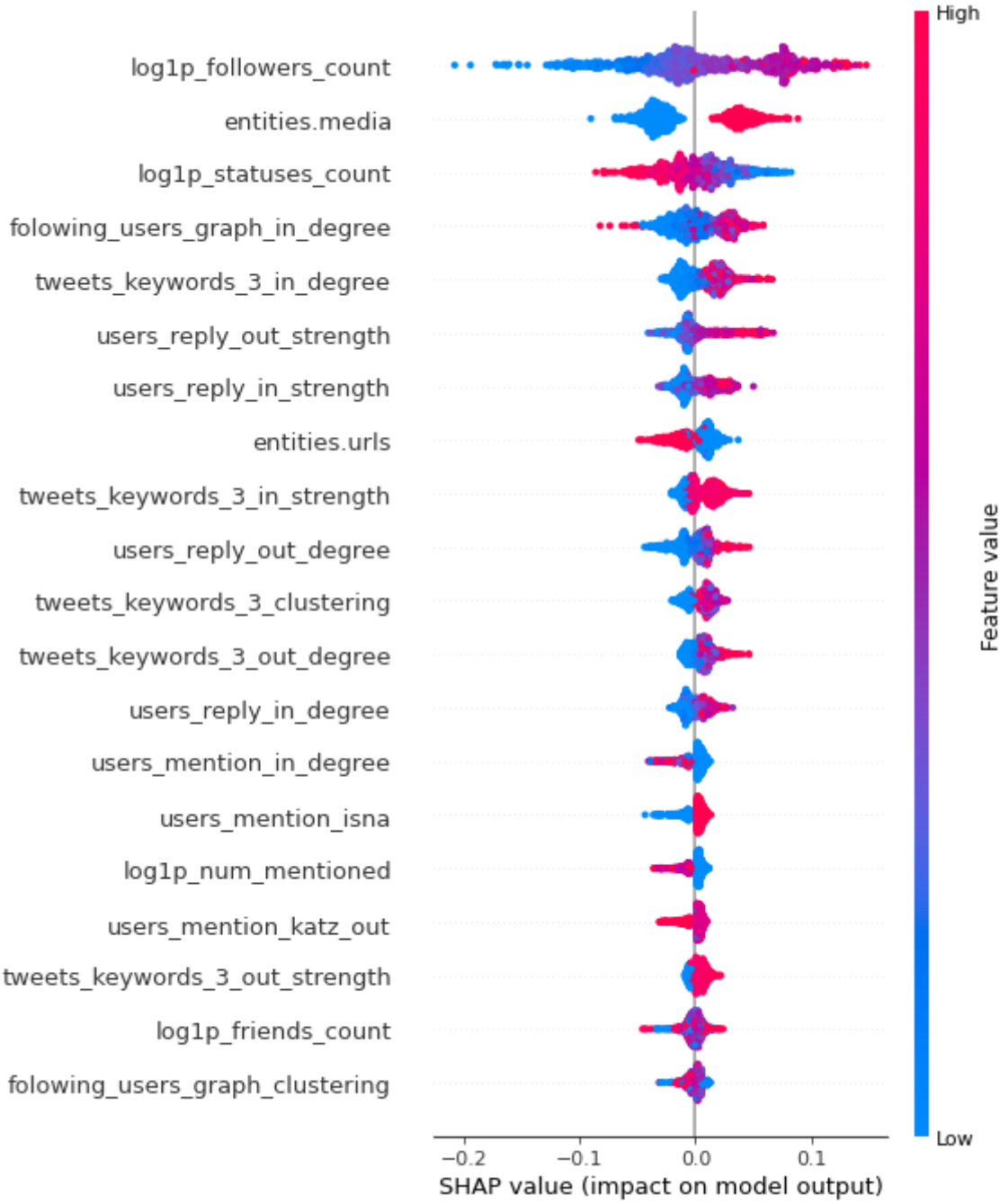
Out[ ]:



Feature importance

In [ ]:

```
# influence on class 1
shap.summary_plot(shap_values.values[:, :, 1], test_df_sample, plot_type='d
```



Feature importance and correlation

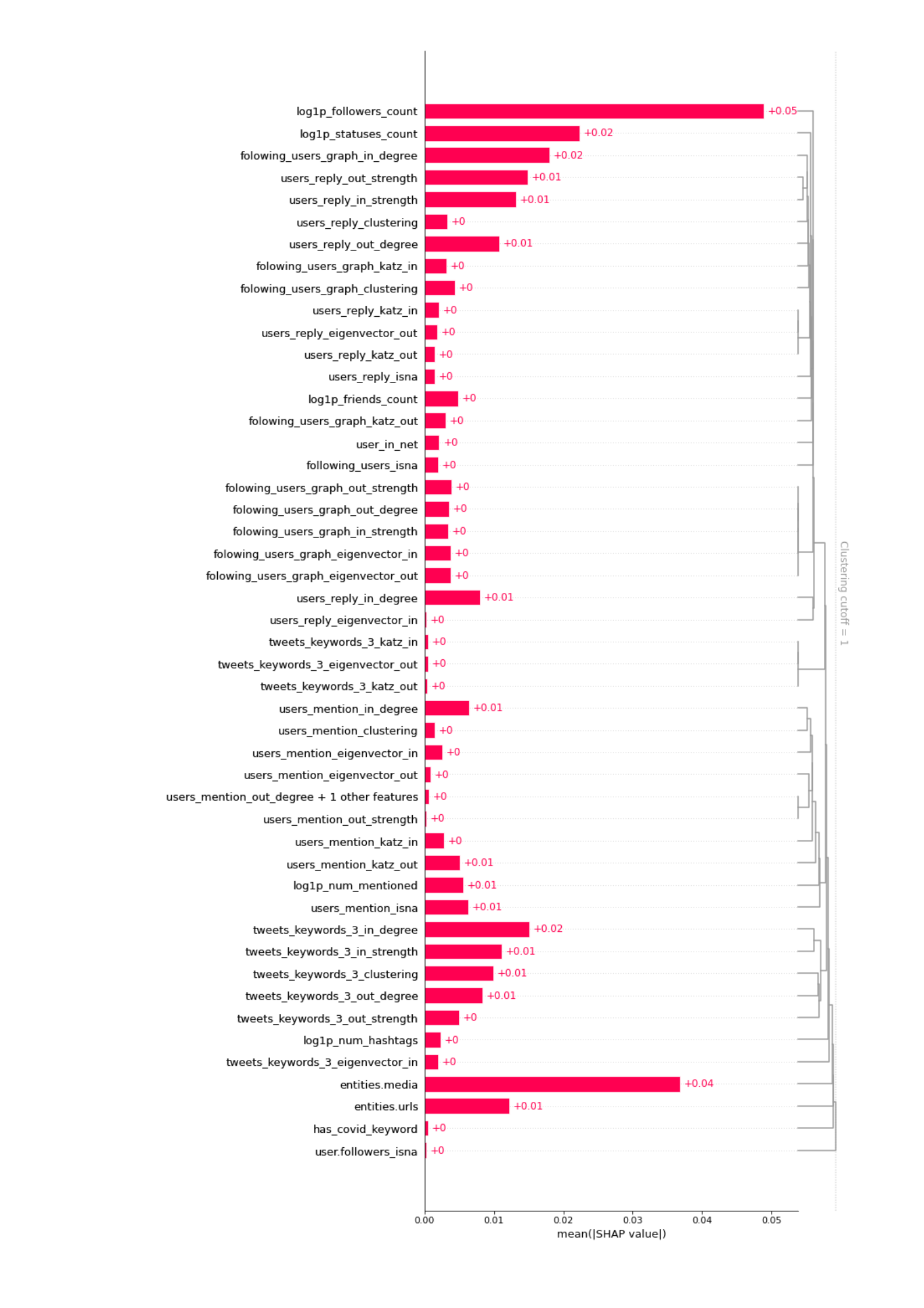
In [ ]:

```
clust = shap.utils.hclust(test_df_sample, test_df_labels_sample, linkage="s
```

```
`early_stopping_rounds` in `fit` method is deprecated for better compatibil
ity with scikit-learn, use `early_stopping_rounds` in constructor or`set_pa
rams` instead.
No/low signal found from feature 2 (this is typically caused by constant or
near-constant features)! Cluster distances can't be computed for it (so set
ting all distances to 1).
No/low signal found from feature 3 (this is typically caused by constant or
near-constant features)! Cluster distances can't be computed for it (so set
ting all distances to 1).
 84%|████████▍  | 41/49 [00:11<00:03,  2.35it/s]No/low signal found from fea
ture 40 (this is typically caused by constant or near-constant features)! C
luster distances can't be computed for it (so setting all distances to 1).
No/low signal found from feature 41 (this is typically caused by constant o
r near-constant features)! Cluster distances can't be computed for it (so s
etting all distances to 1).
No/low signal found from feature 42 (this is typically caused by constant o
r near-constant features)! Cluster distances can't be computed for it (so s
etting all distances to 1).
100%|██████████| 49/49 [00:13<00:00,  3.03it/s]No/low signal found from fea
ture 48 (this is typically caused by constant or near-constant features)! C
luster distances can't be computed for it (so setting all distances to 1).
50it [00:13,  1.04s/it]
```

In [ ]:

```
exp = shap.Explanation(shap_values.values[:, :, 1], shap_values.base_values
shap.plots.bar(exp, max_display=48, clustering=clust, clustering_cutoff=1)
```

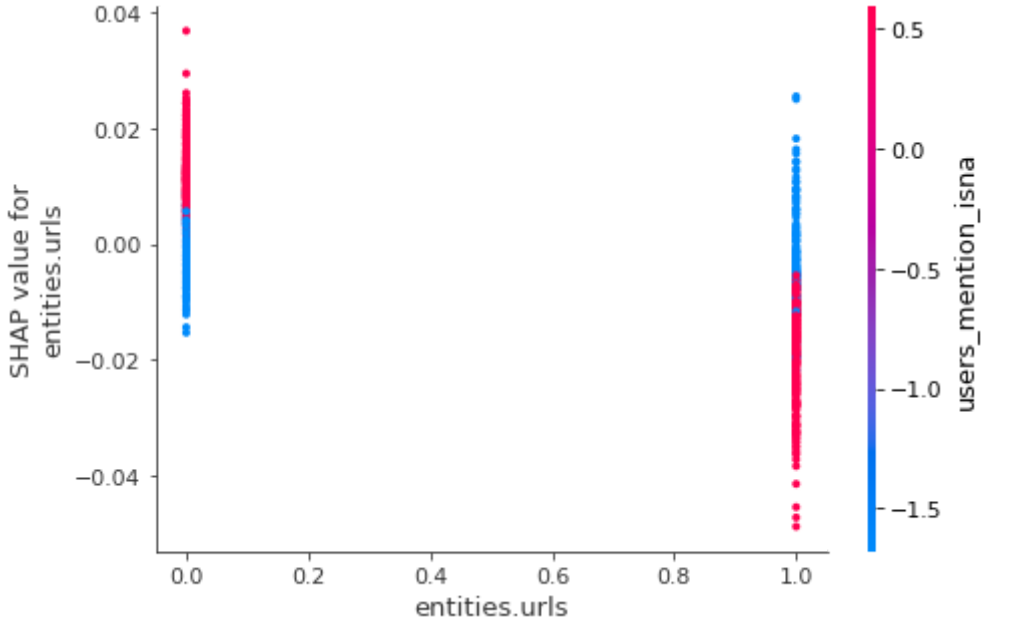

log1p_followers_count +0.05
log1p_statuses_count +0.02
folowing_users_graph_in_degree +0.02
users_reply_out_strength +0.01
users_reply_in_strength +0.01
users_reply_clustering +0
users_reply_out_degree +0.01
folowing_users_graph_katz_in +0
folowing_users_graph_clustering +0
users_reply_katz_in +0
users_reply_eigenvector_out +0
users_reply_katz_out +0
users_reply_isna +0
log1p_friends_count +0
folowing_users_graph_katz_out +0
user_in_net +0
following_users_isna +0
folowing_users_graph_out_strength +0
folowing_users_graph_out_degree +0
folowing_users_graph_in_strength +0
folowing_users_graph_eigenvector_in +0
folowing_users_graph_eigenvector_out +0
users_reply_in_degree +0.01
users_reply_eigenvector_in +0
tweets_keywords_3_katz_in +0
tweets_keywords_3_eigenvector_out +0
tweets_keywords_3_katz_out +0
users_mention_in_degree +0.01
users_mention_clustering +0
users_mention_eigenvector_in +0
users_mention_eigenvector_out +0
users_mention_out_degree + 1 other features +0
users_mention_out_strength +0
users_mention_katz_in +0
users_mention_katz_out +0.01
log1p_num_mentioned +0.01
users_mention_isna +0.01
tweets_keywords_3_in_degree +0.02
tweets_keywords_3_in_strength +0.01
tweets_keywords_3_clustering +0.01
tweets_keywords_3_out_degree +0.01
tweets_keywords_3_out_strength +0
log1p_num_hashtags +0
tweets_keywords_3_eigenvector_in +0
entities.media +0.04
entities.urls +0.01
has_covid_keyword +0
user.followers_isna +0
0.00
0.01
0.02
0.03
0.04
0.05
mean(|SHAP value|)
Clustering cutoff = 1

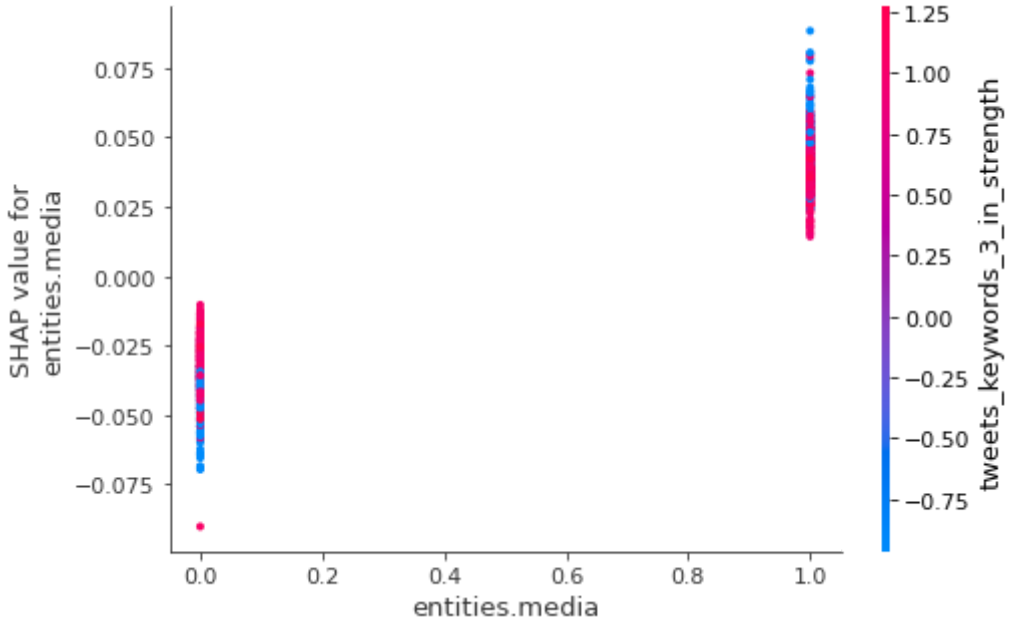
SHAP dependence plots

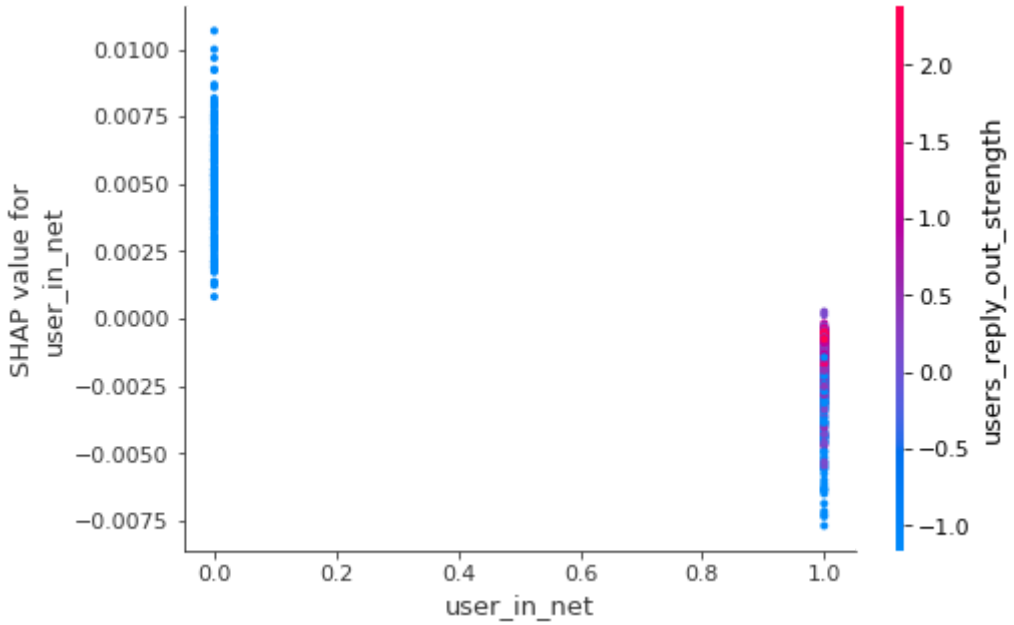
"SHAP dependence plots show the effect of a single feature across the whole dataset. They plot a feature's value vs. the SHAP value of that feature across many samples. SHAP dependence plots are similar to partial dependence plots, but account for the interaction effects present in the features, and are only defined in regions of the input space supported by data. The vertical dispersion of SHAP values at a single feature value is driven by interaction effects, and another feature is chosen for coloring to highlight possible interactions."

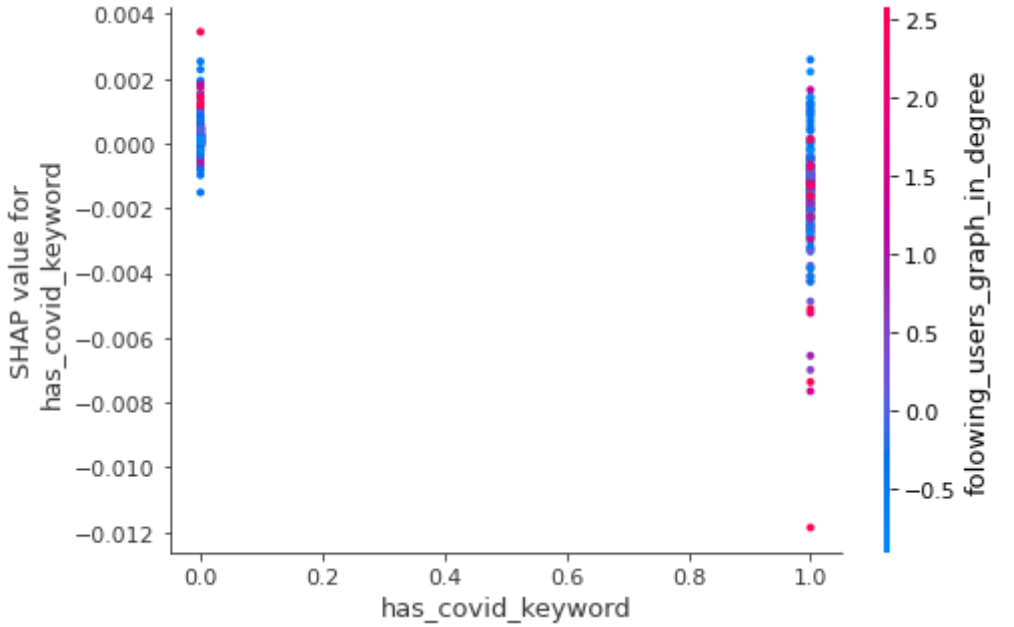
https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/

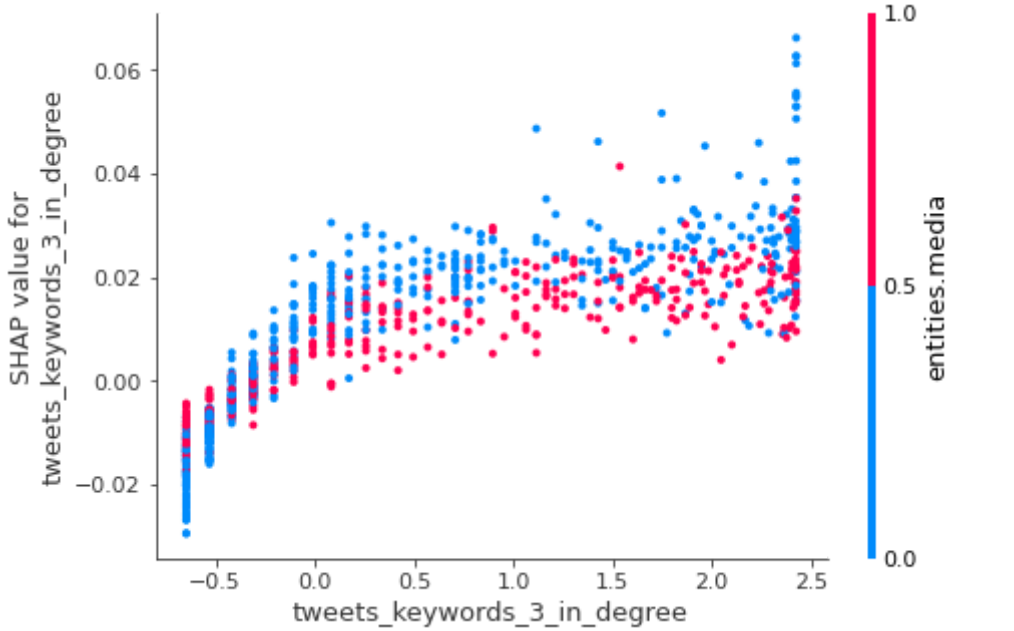
In [ ]:
```
for name in test_df_sample.columns:
    shap.dependence_plot(name, shap_values.values[:, :, 1], test_df_sample,
```

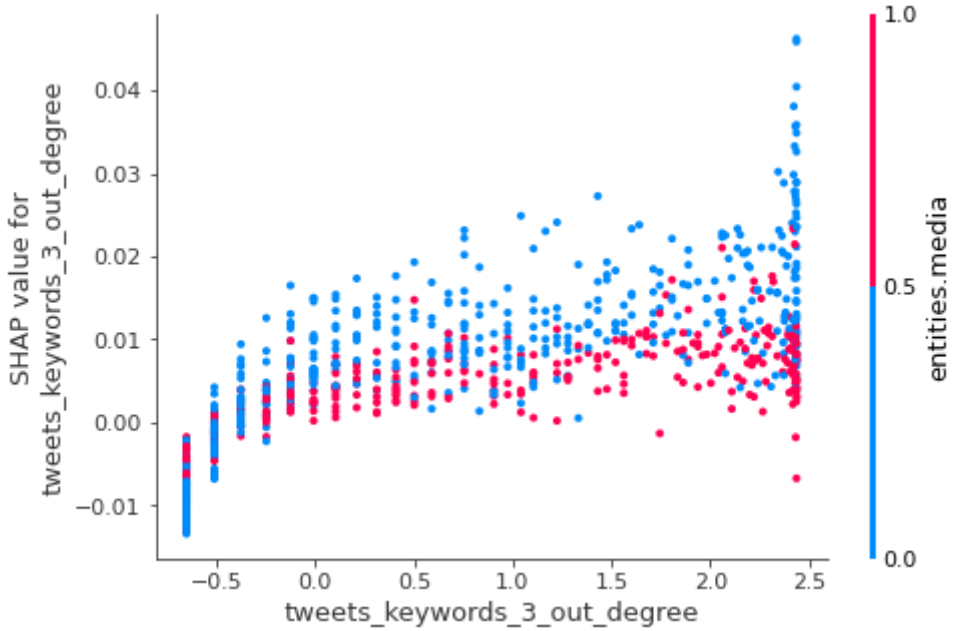


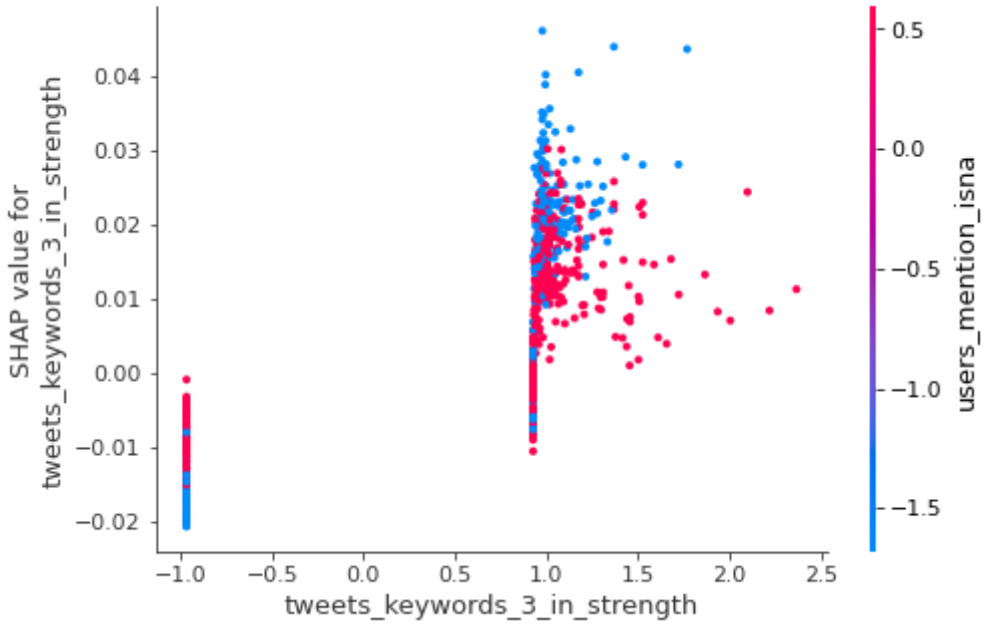


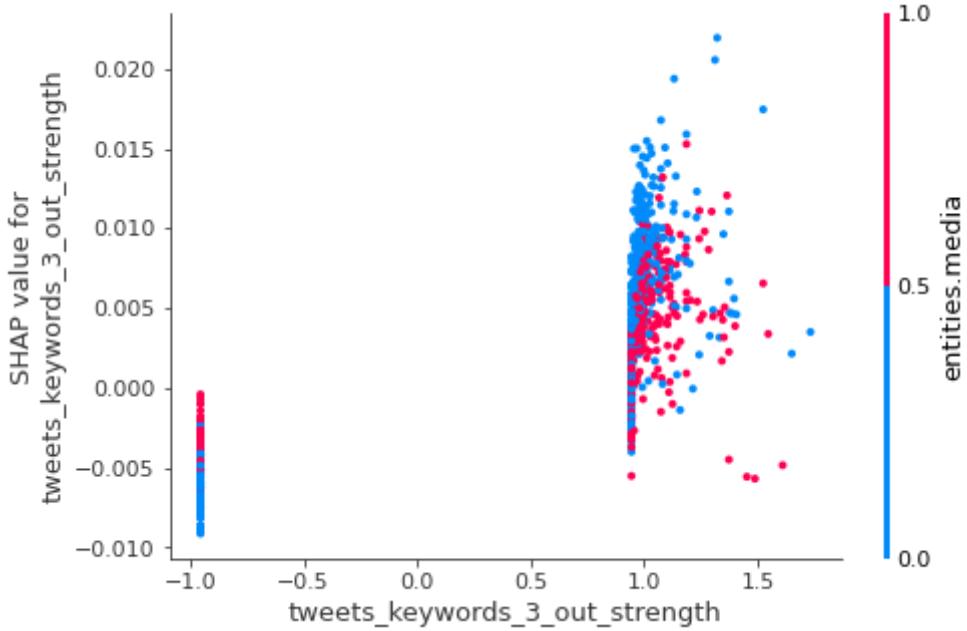

SHAP value for
user_in_net
0.0100
0.0075
0.0050
0.0025
0.0000
−0.0025
−0.0050
−0.0075
0.0
0.2
0.4
0.6
0.8
1.0
user_in_net
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
users_reply_out_strength

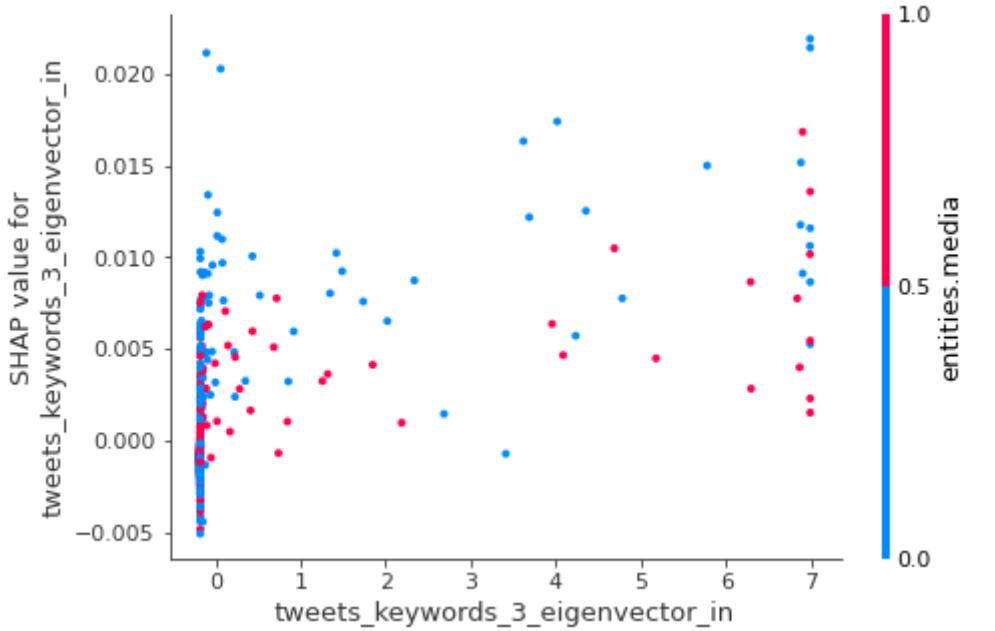

SHAP value for
has_covid_keyword
0.004
0.002
0.000
−0.002
−0.004
−0.006
−0.008
−0.010
−0.012
0.0
0.2
0.4
0.6
0.8
1.0
has_covid_keyword
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
folowing_users_graph_in_degree

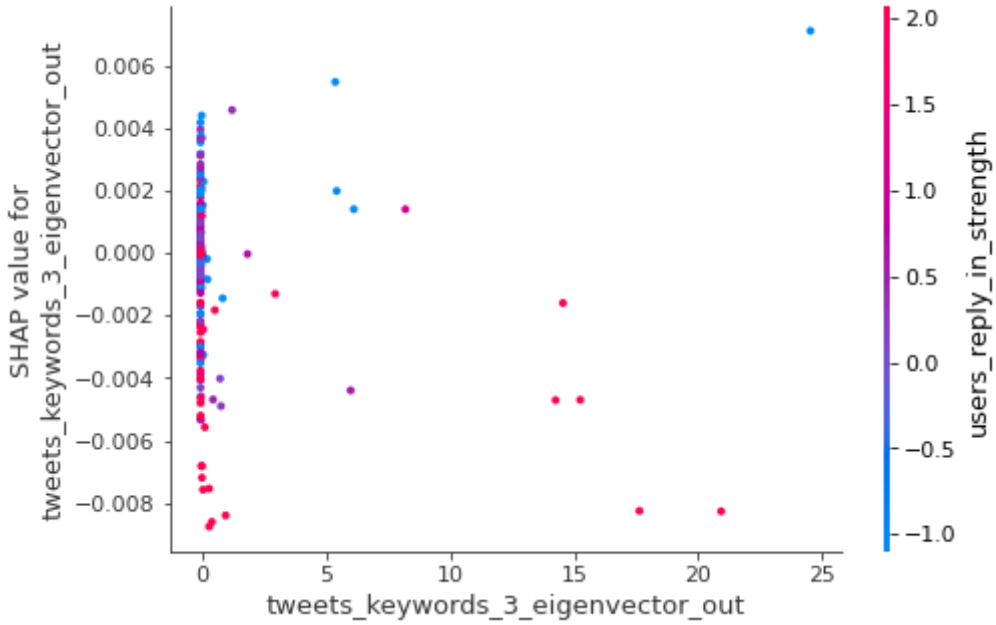

SHAP value for
tweets_keywords_3_in_degree
0.06
0.04
0.02
0.00
−0.02
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
tweets_keywords_3_in_degree
1.0
0.5
0.0
entities.media

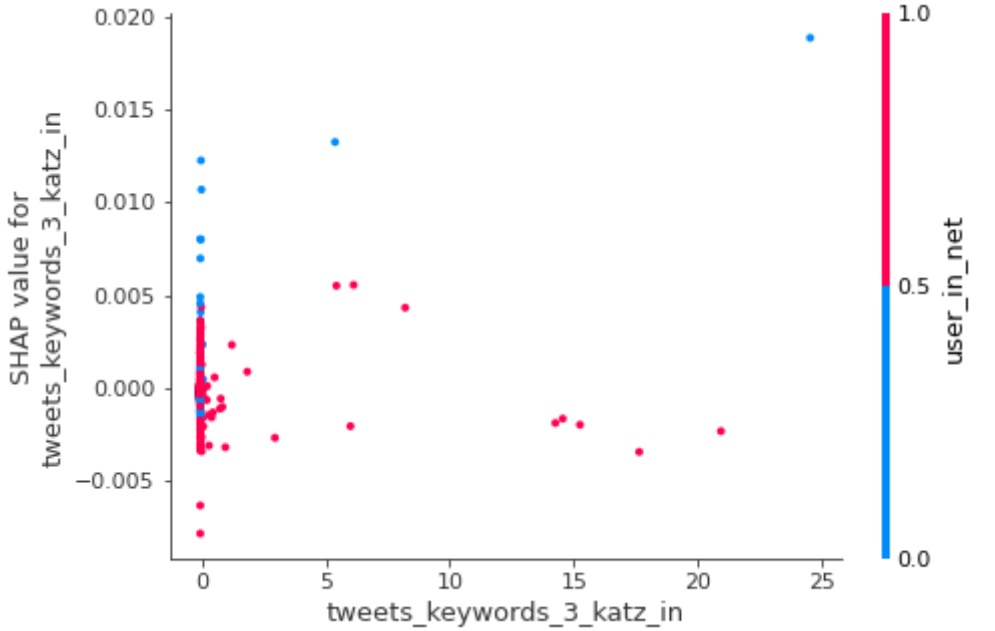

SHAP value for
tweets_keywords_3_out_degree
0.04
0.03
0.02
0.01
0.00
−0.01
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
tweets_keywords_3_out_degree
1.0
0.5
0.0
entities.media

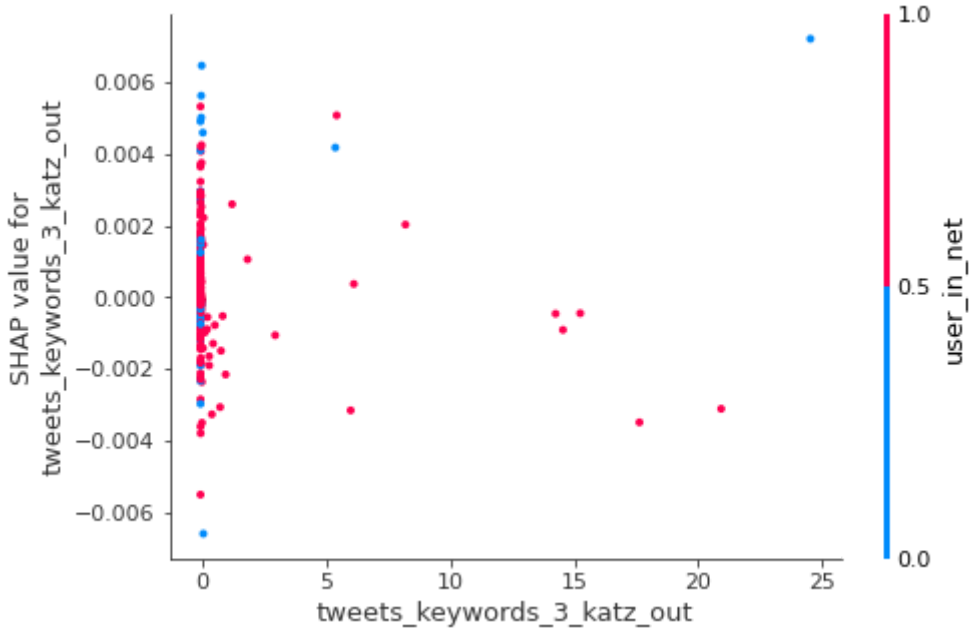

SHAP value for
tweets_keywords_3_in_strength
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
tweets_keywords_3_in_strength
0.5
0.0
−0.5
−1.0
−1.5
users_mention_isna

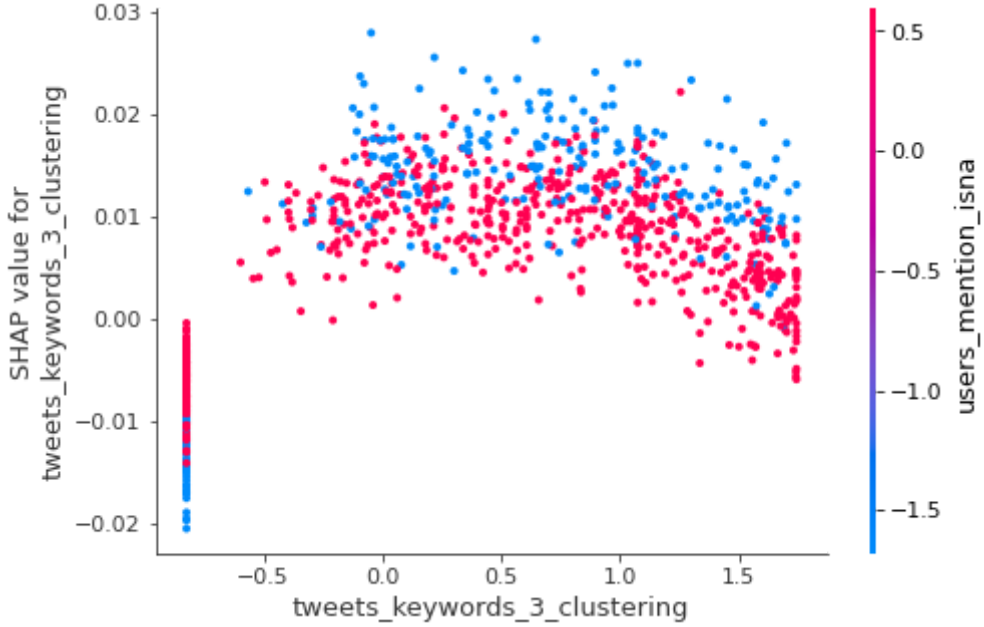

SHAP value for
tweets_keywords_3_out_strength
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010
−1.0
−0.5
0.0
0.5
1.0
1.5
tweets_keywords_3_out_strength
1.0
0.5
0.0
entities.media

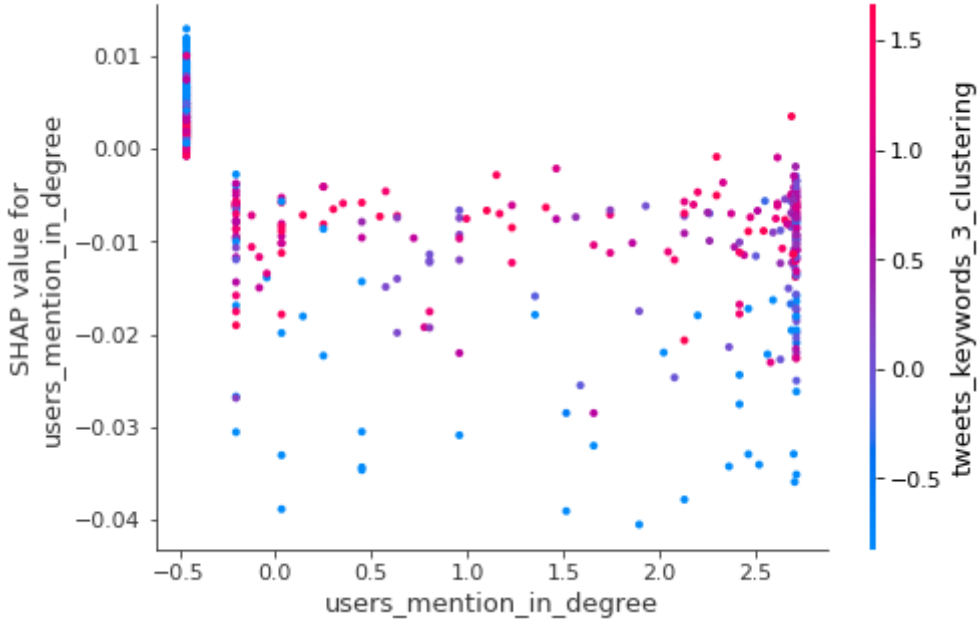

SHAP value for
tweets_keywords_3_eigenvector_in
0.020
0.015
0.010
0.005
0.000
−0.005
0
1
2
3
4
5
6
7
tweets_keywords_3_eigenvector_in
1.0
0.5
0.0
entities.media

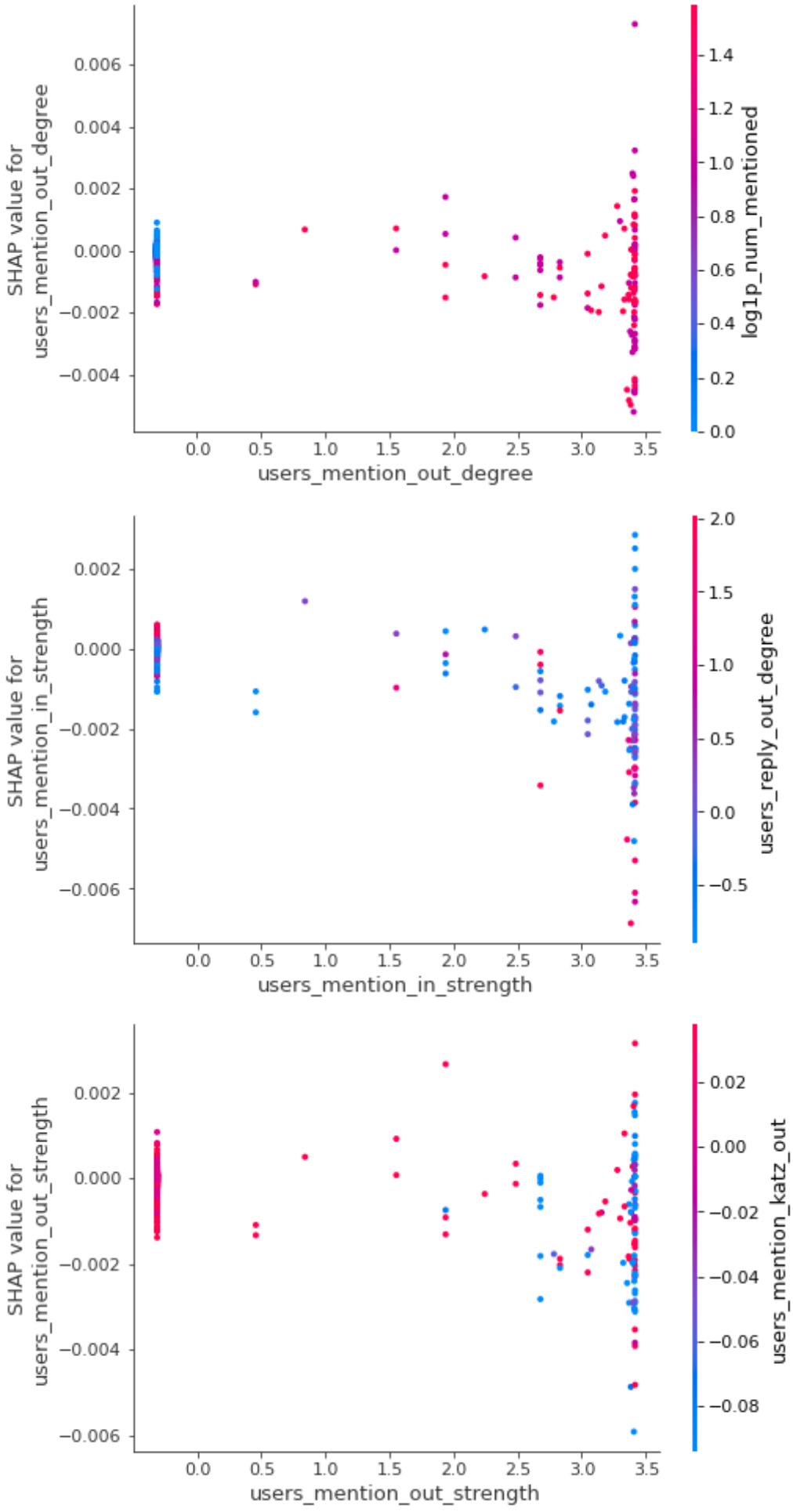

SHAP value for
tweets_keywords_3_eigenvector_out
0.006
0.004
0.002
0.000
−0.002
−0.004
−0.006
−0.008
0
5
10
15
20
25
tweets_keywords_3_eigenvector_out
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
users_reply_in_strength

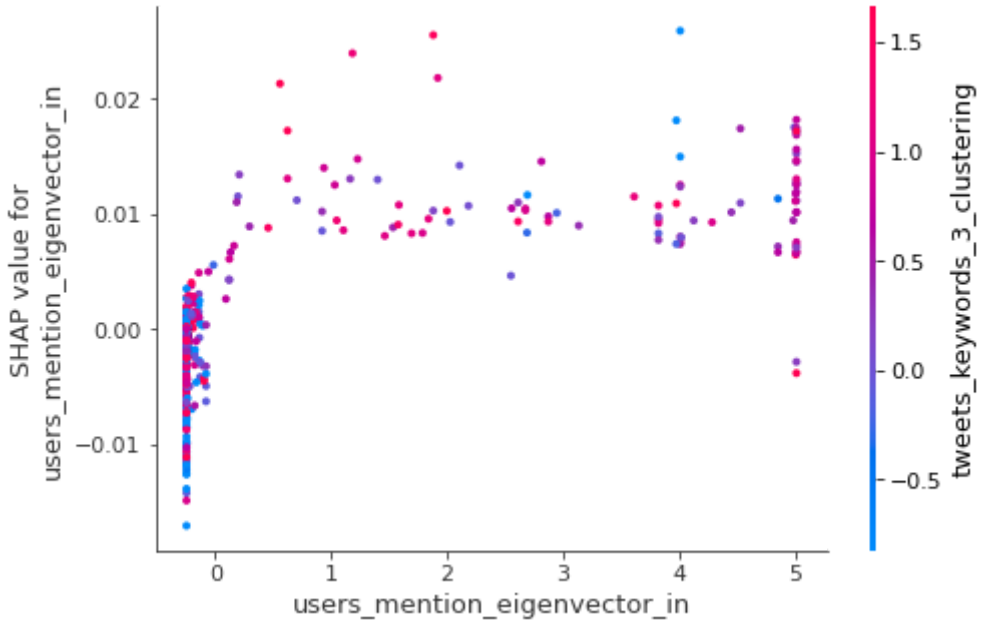

SHAP value for
tweets_keywords_3_katz_in
0.020
0.015
0.010
0.005
0.000
−0.005
0
5
10
15
20
25
tweets_keywords_3_katz_in
1.0
0.5
0.0
user_in_net

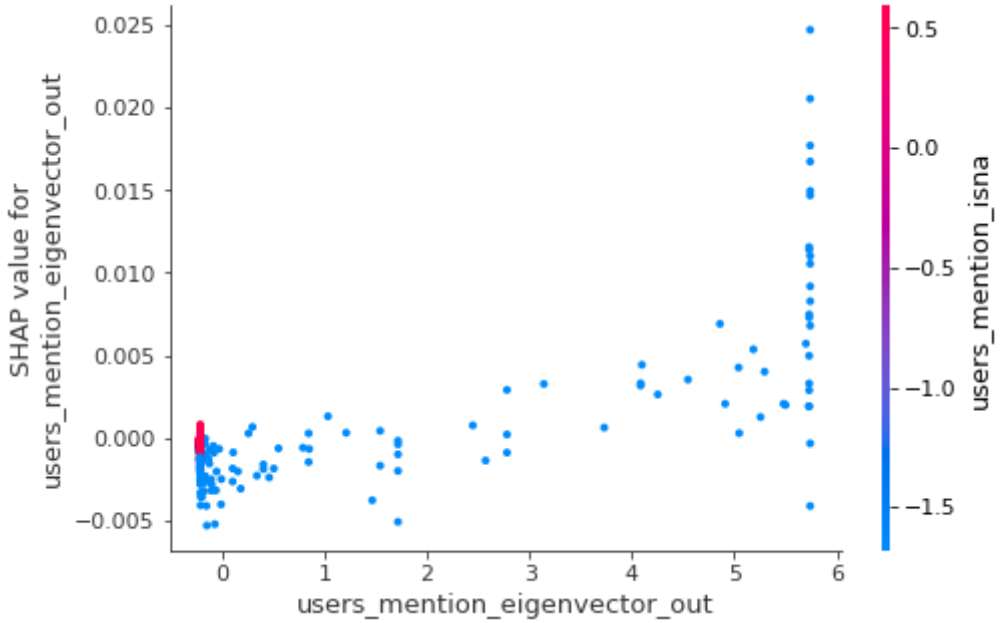

SHAP value for
tweets_keywords_3_katz_out
0.006
0.004
0.002
0.000
−0.002
−0.004
−0.006
0
5
10
15
20
25
tweets_keywords_3_katz_out
1.0
0.5
0.0
user_in_net

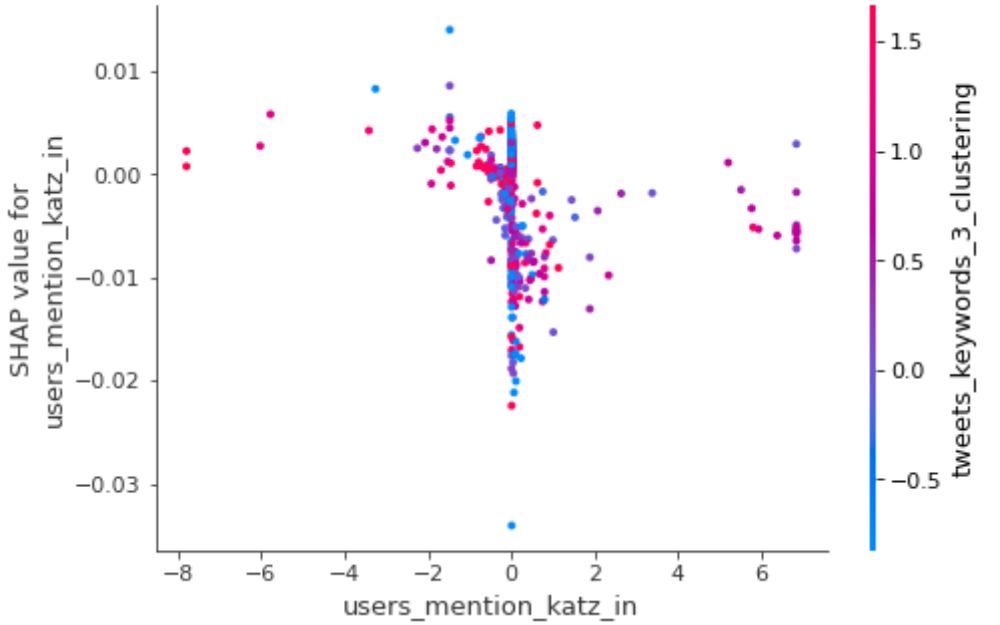

SHAP value for
tweets_keywords_3_clustering
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.5
0.0
0.5
1.0
1.5
tweets_keywords_3_clustering
0.5
0.0
−0.5
−1.0
−1.5
users_mention_isna

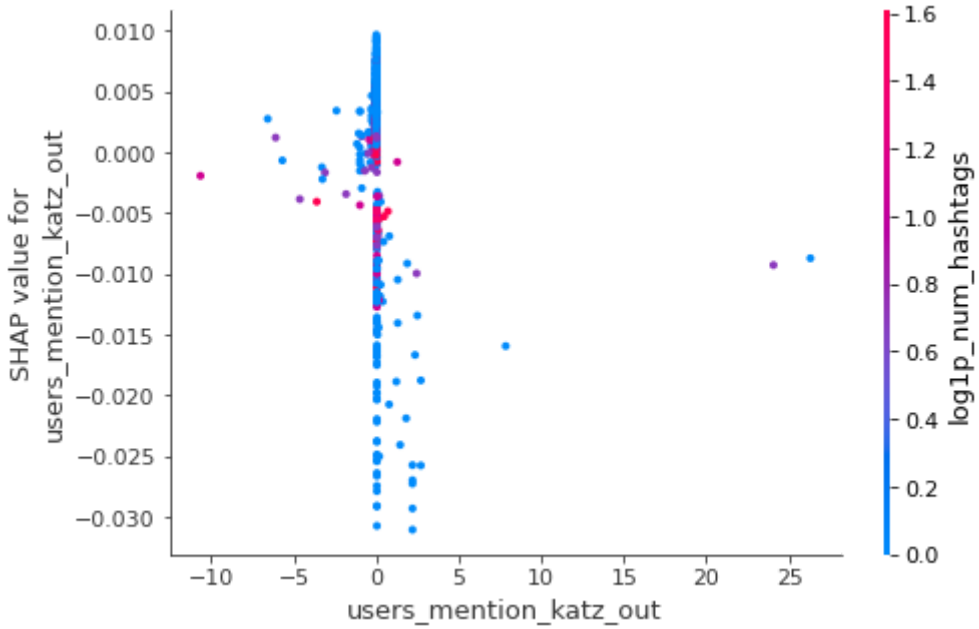

SHAP value for
users_mention_in_degree
0.01
0.00
−0.01
−0.02
−0.03
−0.04
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
users_mention_in_degree
1.5
1.0
0.5
0.0
−0.5
tweets_keywords_3_clustering

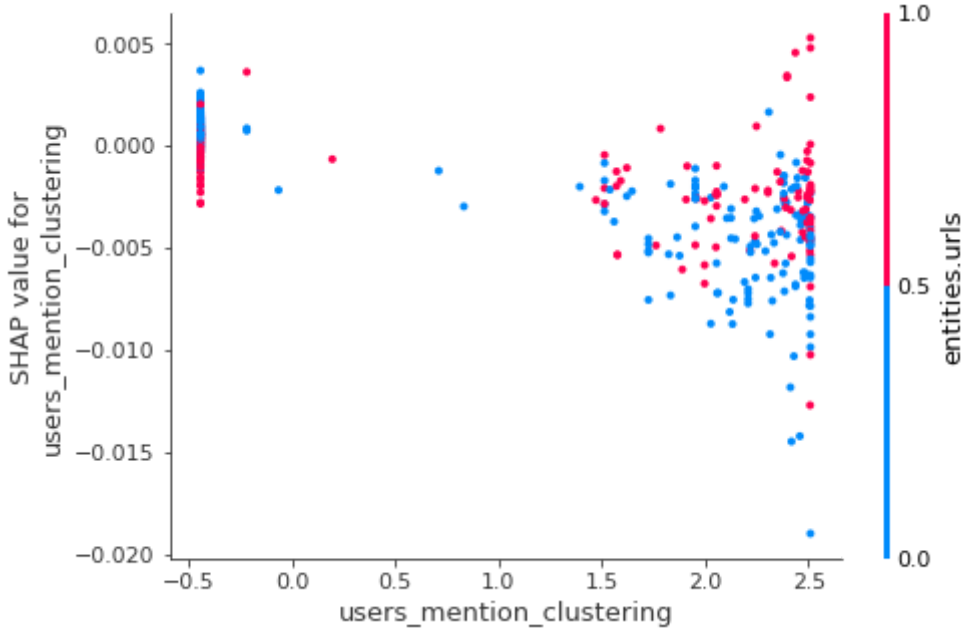

SHAP value for
users_mention_out_degree
0.006
0.004
0.002
0.000
−0.002
−0.004
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
users_mention_out_degree
log1p_num_mentioned
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
SHAP value for
users_mention_in_strength
0.002
0.000
−0.002
−0.004
−0.006
users_mention_in_strength
users_reply_out_degree
2.0
1.5
1.0
0.5
0.0
−0.5
SHAP value for
users_mention_out_strength
0.002
0.000
−0.002
−0.004
−0.006
users_mention_out_strength
users_mention_katz_out
0.02
0.00
−0.02
−0.04
−0.06
−0.08

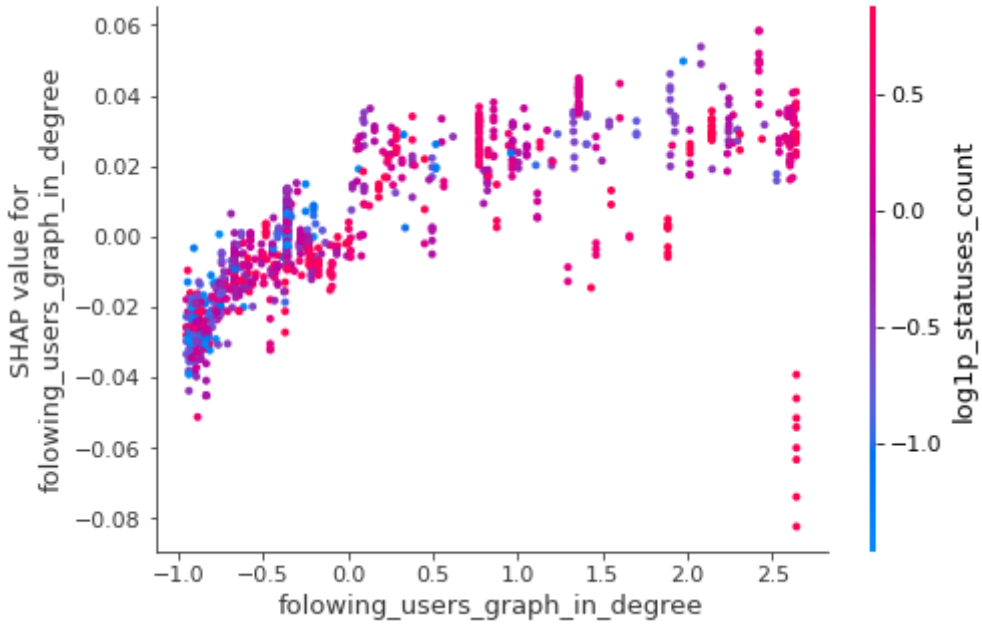

SHAP value for
users_mention_eigenvector_in
users_mention_eigenvector_in
tweets_keywords_3_clustering

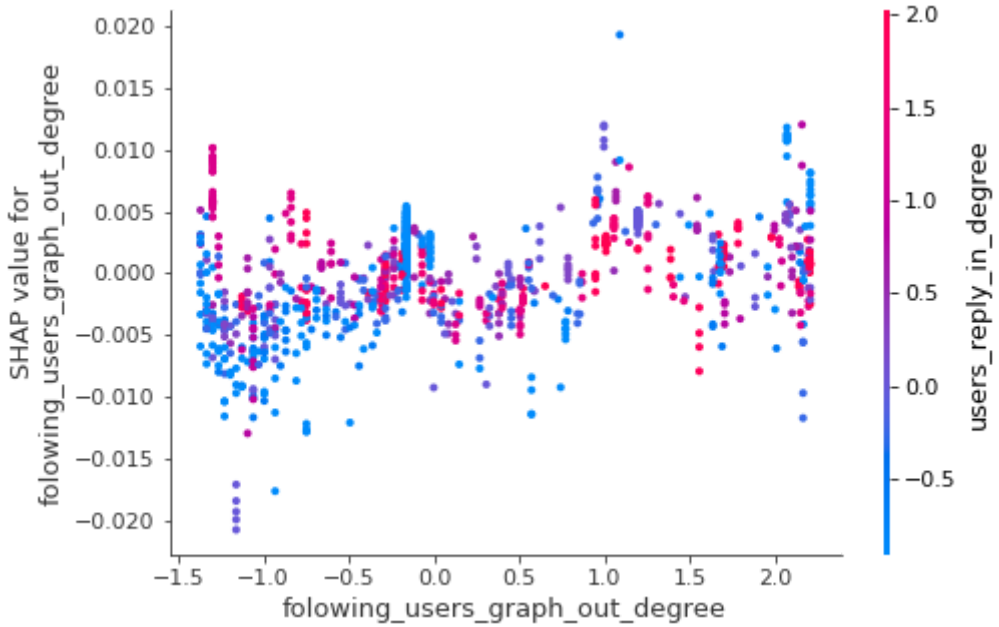

SHAP value for
users_mention_eigenvector_out
users_mention_eigenvector_out
users_mention_isna

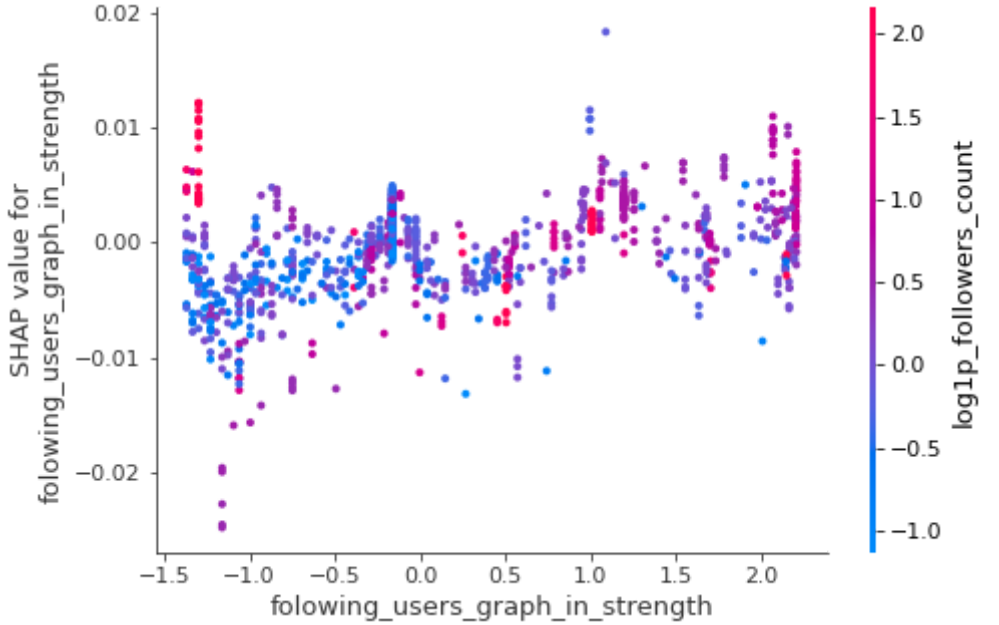

SHAP value for
users_mention_katz_in
users_mention_katz_in
tweets_keywords_3_clustering

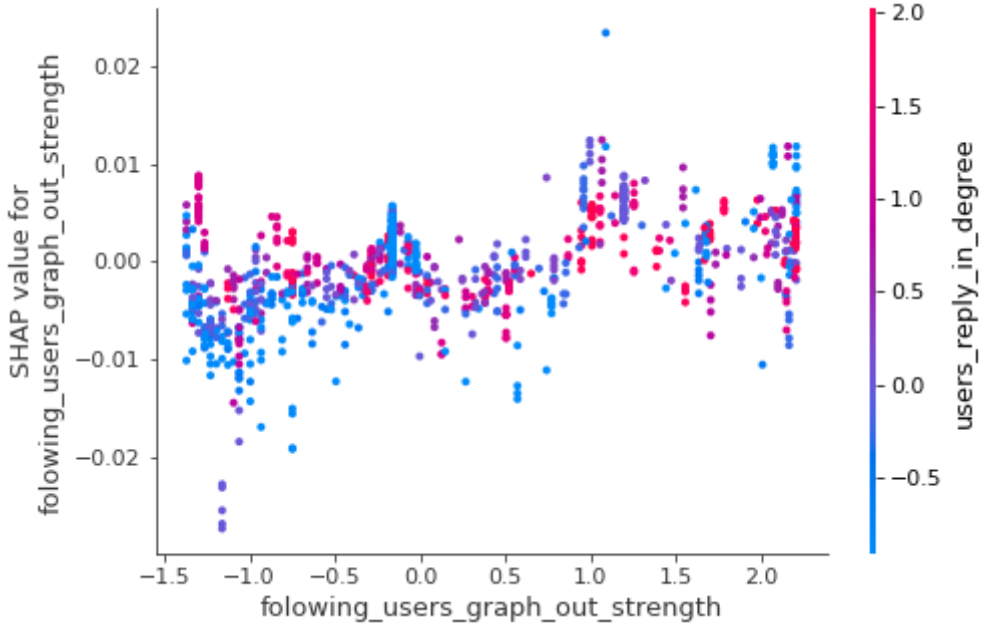

SHAP value for
users_mention_katz_out
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
−0.025
−0.030
−10
−5
0
5
10
15
20
25
users_mention_katz_out
log1p_num_hashtags
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0

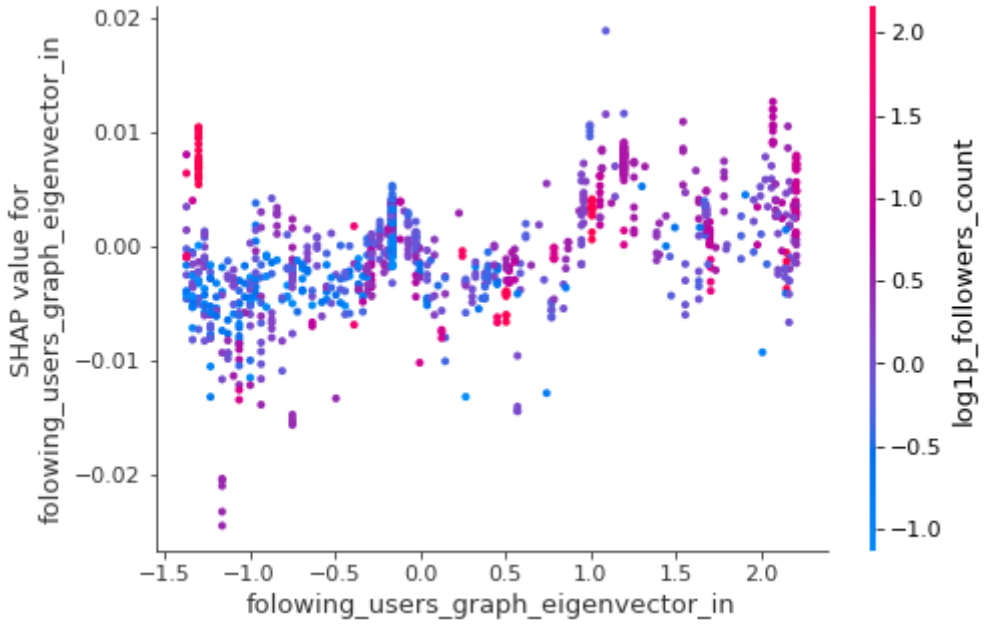

SHAP value for
users_mention_clustering
0.005
0.000
−0.005
−0.010
−0.015
−0.020
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
users_mention_clustering
entities.urls
1.0
0.5
0.0

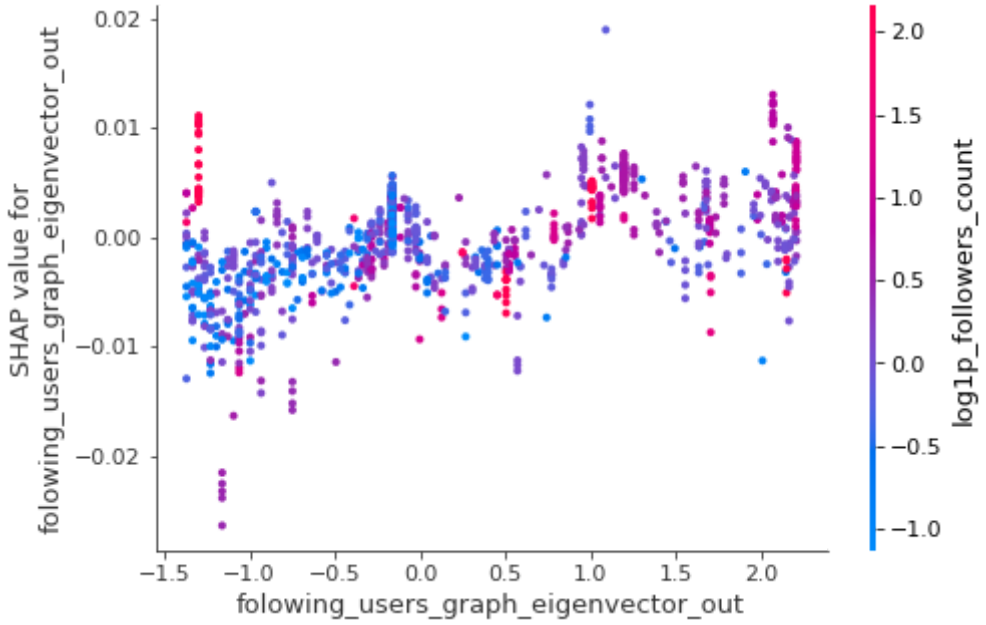

SHAP value for
folowing_users_graph_in_degree
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
folowing_users_graph_in_degree
log1p_statuses_count
0.5
0.0
−0.5
−1.0

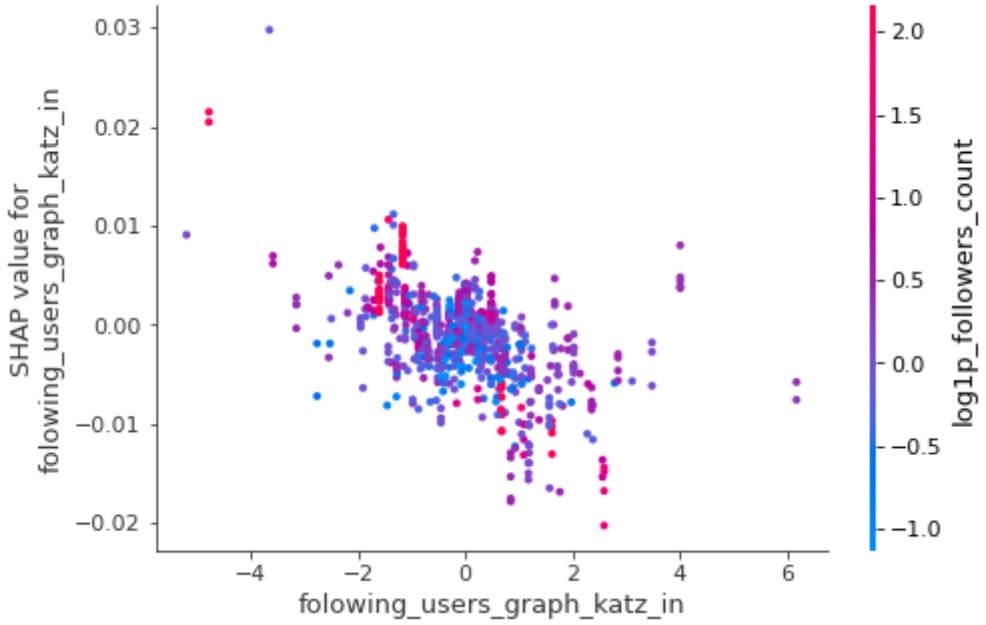

SHAP value for folowing_users_graph_out_degree
folowing_users_graph_out_degree
users_reply_in_degree

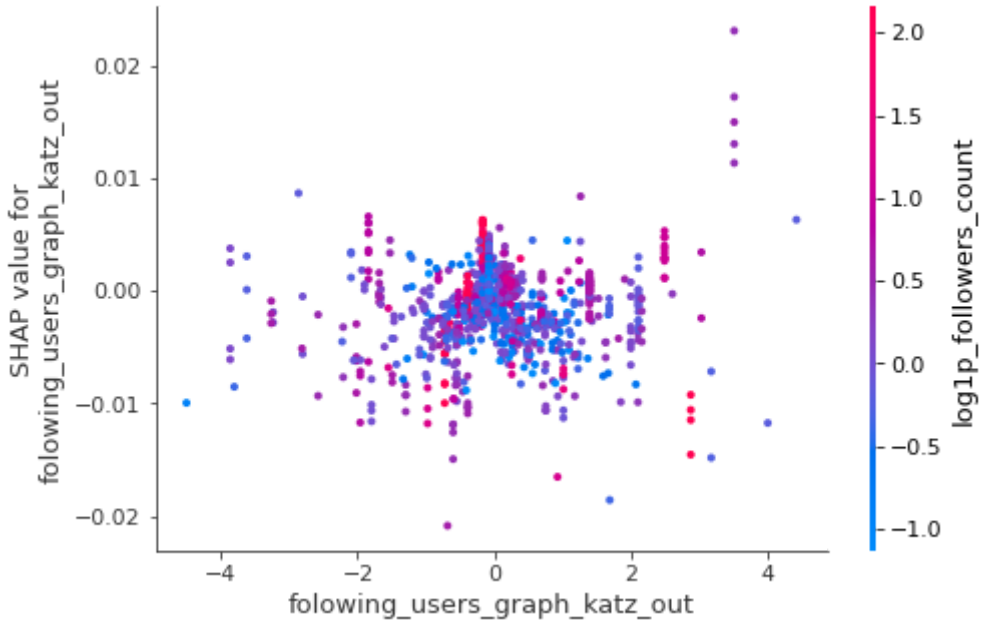

SHAP value for folowing_users_graph_in_strength
folowing_users_graph_in_strength
log1p_followers_count

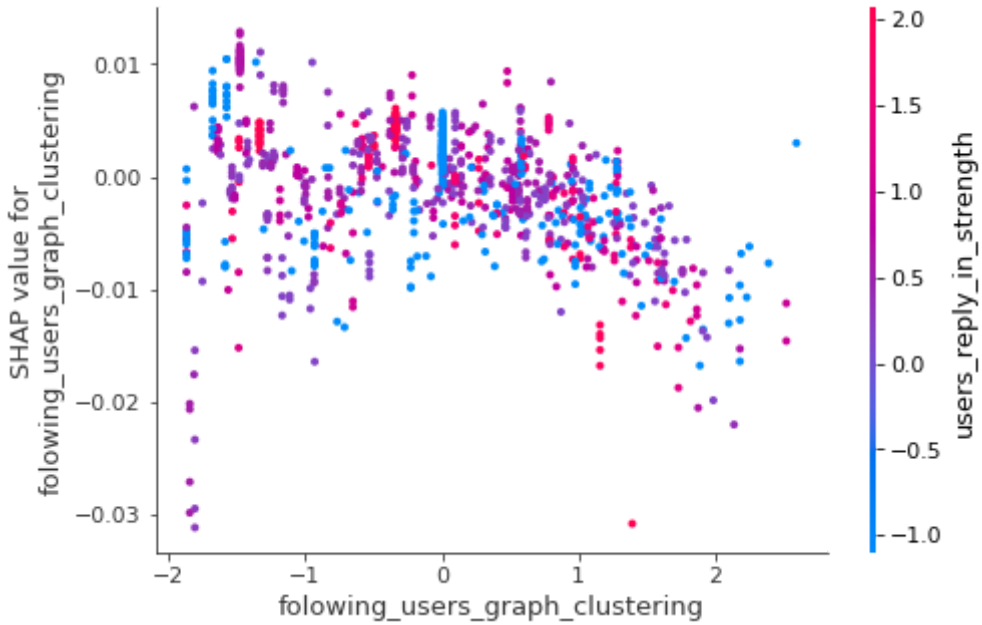

SHAP value for folowing_users_graph_out_strength
folowing_users_graph_out_strength
users_reply_in_degree

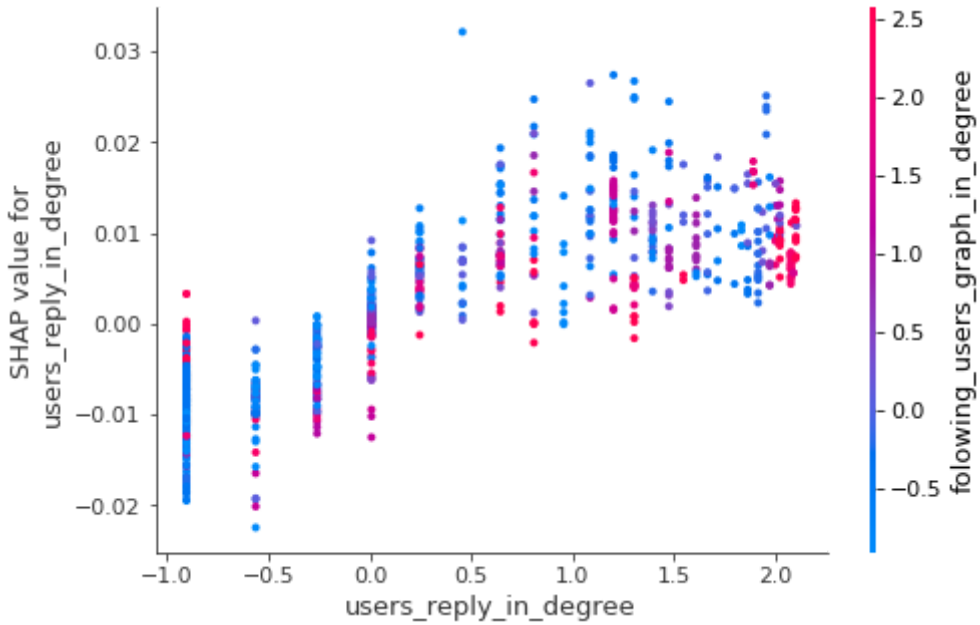

SHAP value for folowing_users_graph_eigenvector_in
folowing_users_graph_eigenvector_in
log1p_followers_count

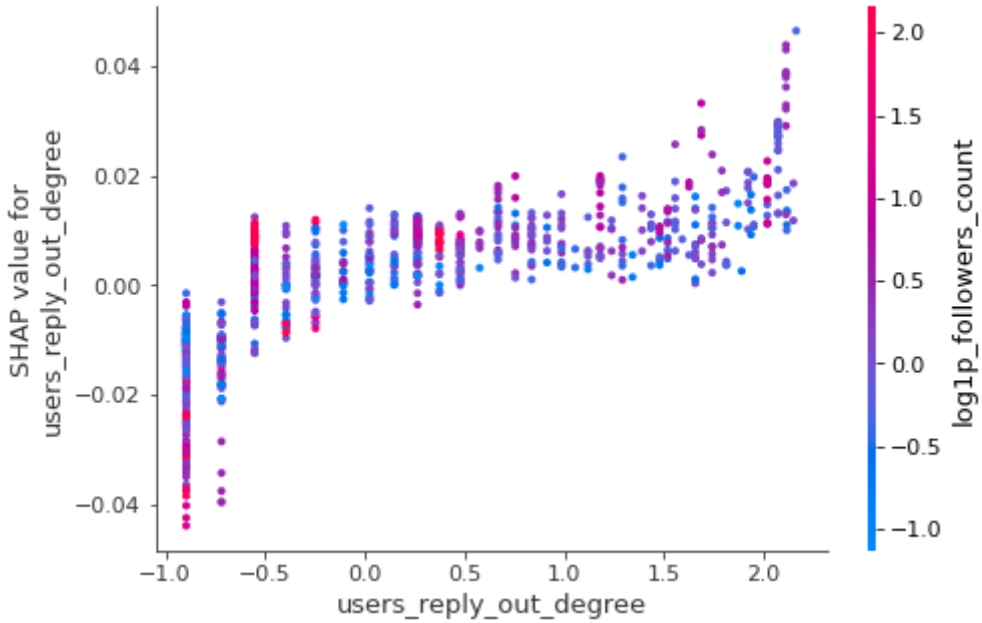

SHAP value for folowing_users_graph_eigenvector_out
folowing_users_graph_eigenvector_out
log1p_followers_count

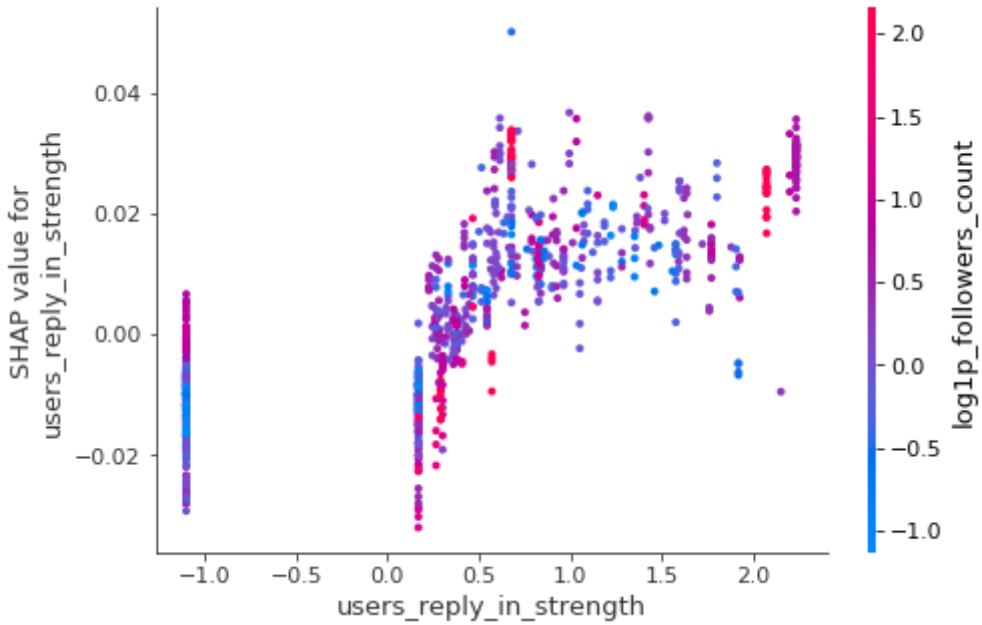

SHAP value for folowing_users_graph_katz_in
folowing_users_graph_katz_in
log1p_followers_count

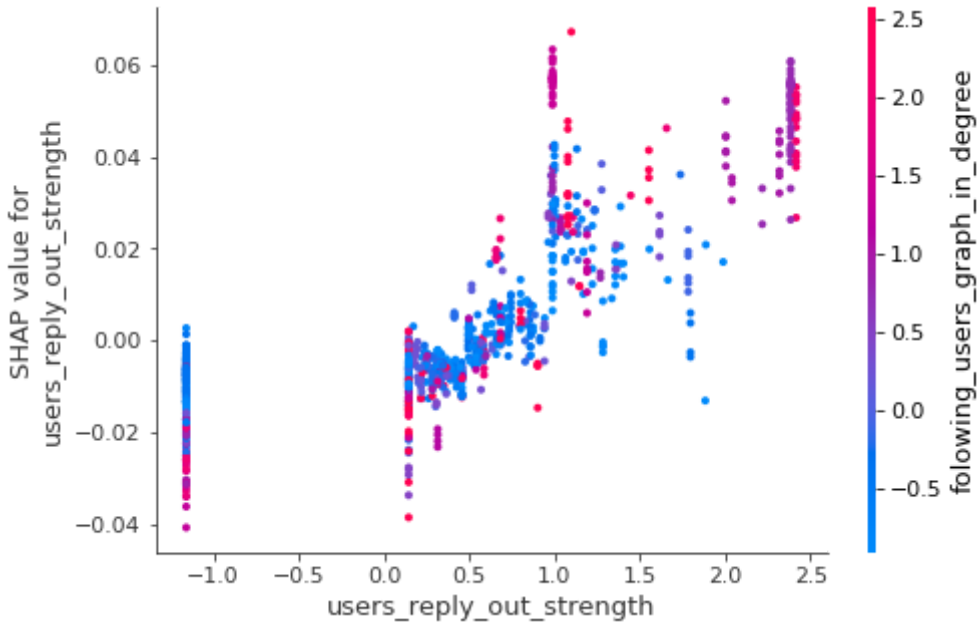

SHAP value for
folowing_users_graph_katz_out
folowing_users_graph_katz_out
log1p_followers_count

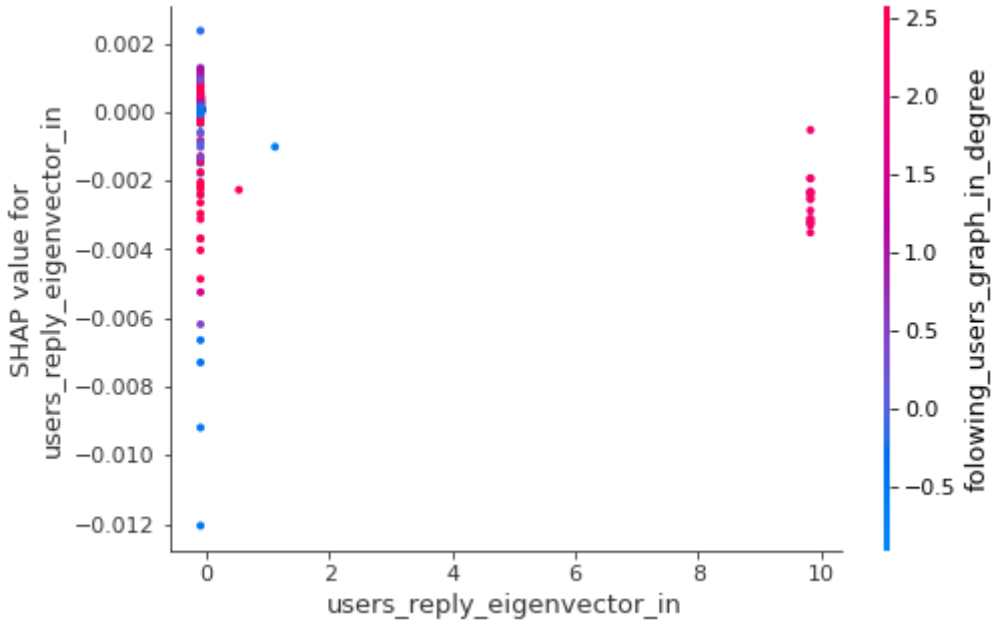

SHAP value for
folowing_users_graph_clustering
folowing_users_graph_clustering
users_reply_in_strength

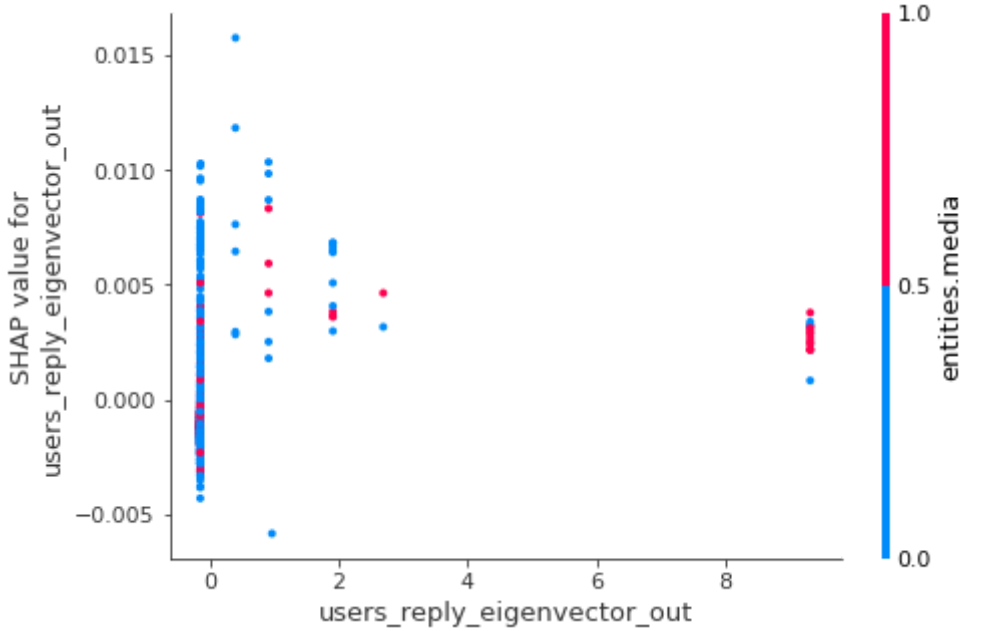

SHAP value for
users_reply_in_degree
users_reply_in_degree
folowing_users_graph_in_degree

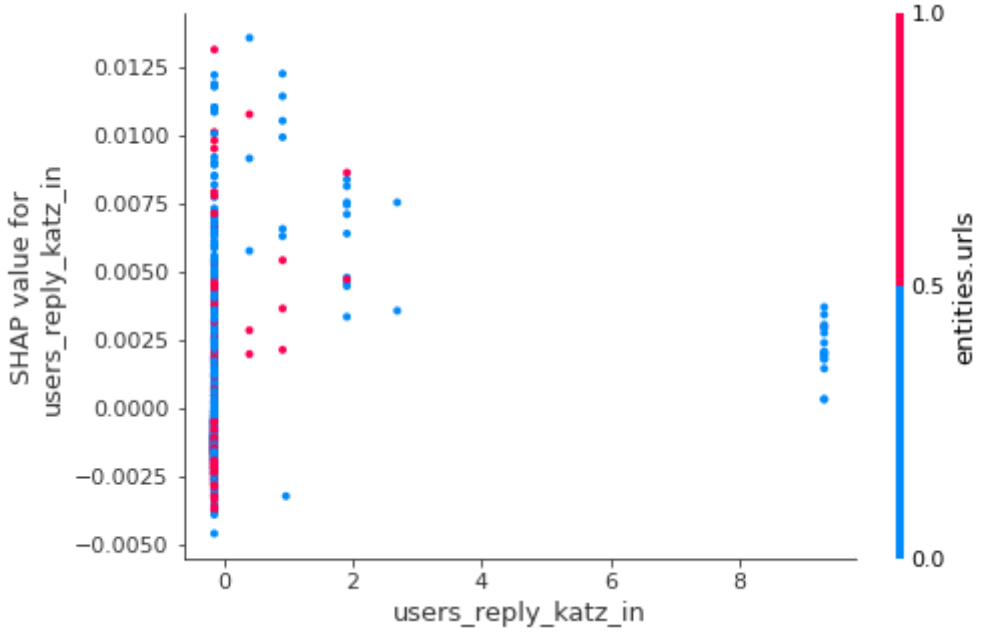

SHAP value for
users_reply_out_degree
users_reply_out_degree
log1p_followers_count

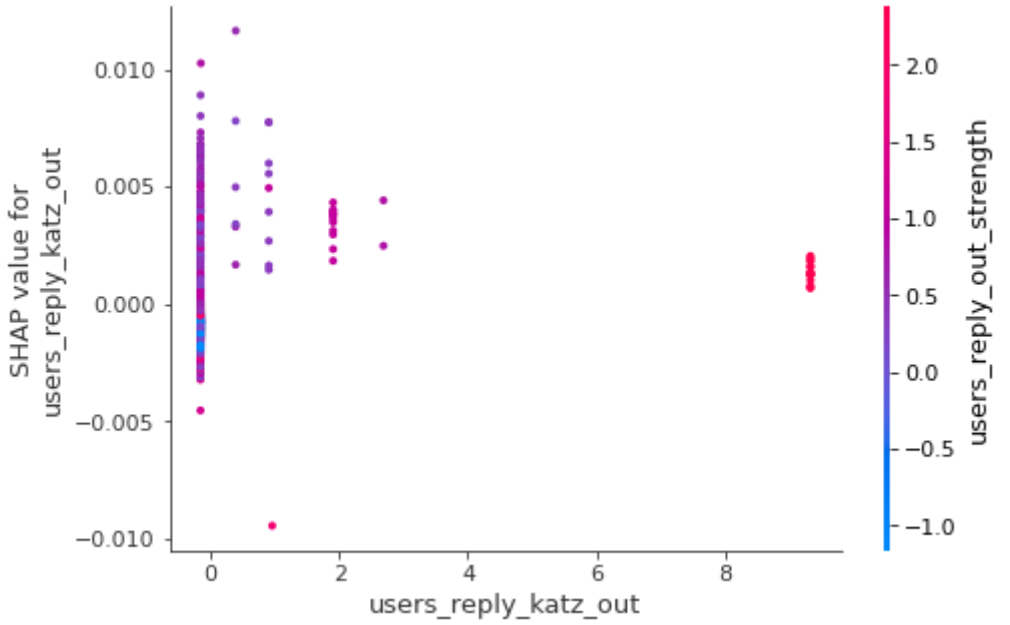

SHAP value for
users_reply_in_strength
users_reply_in_strength
log1p_followers_count

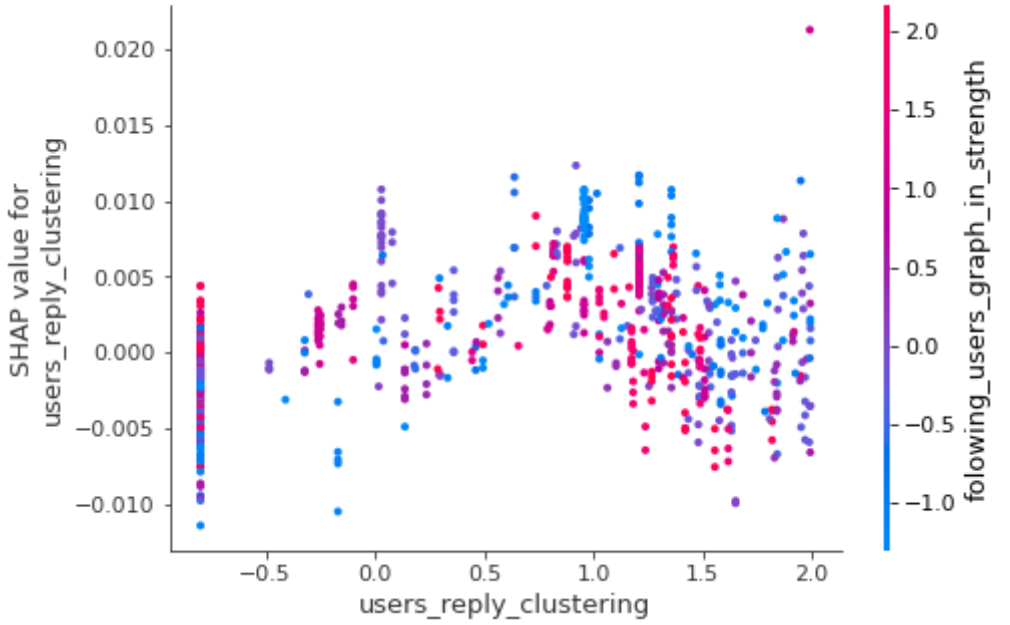

SHAP value for
users_reply_out_strength
users_reply_out_strength
folowing_users_graph_in_degree

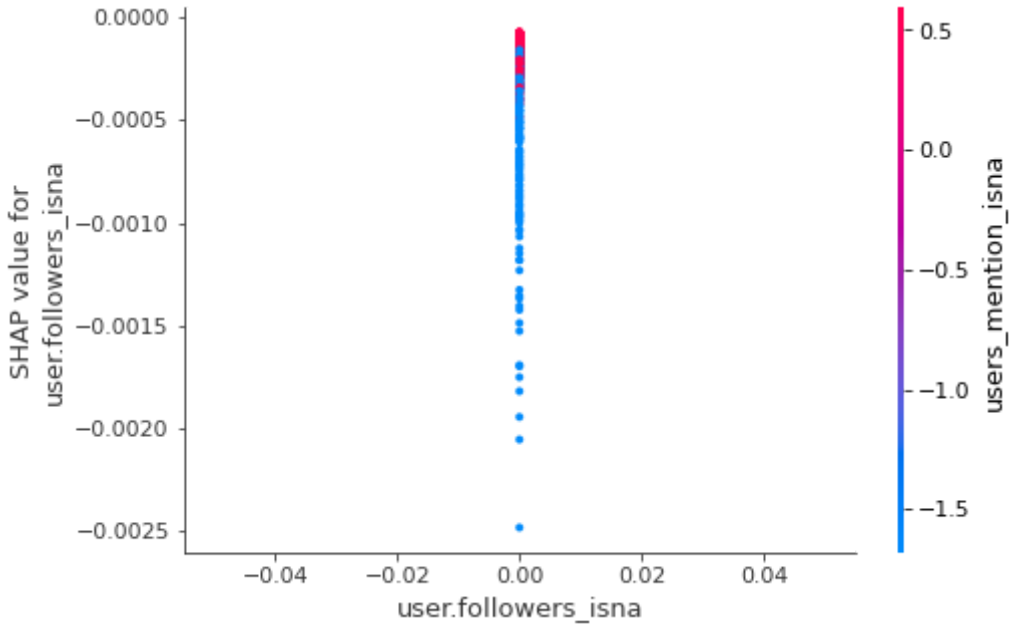

SHAP value for
users_reply_eigenvector_in
0.002
0.000
−0.002
−0.004
−0.006
−0.008
−0.010
−0.012
0
2
4
6
8
10
users_reply_eigenvector_in
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
folowing_users_graph_in_degree

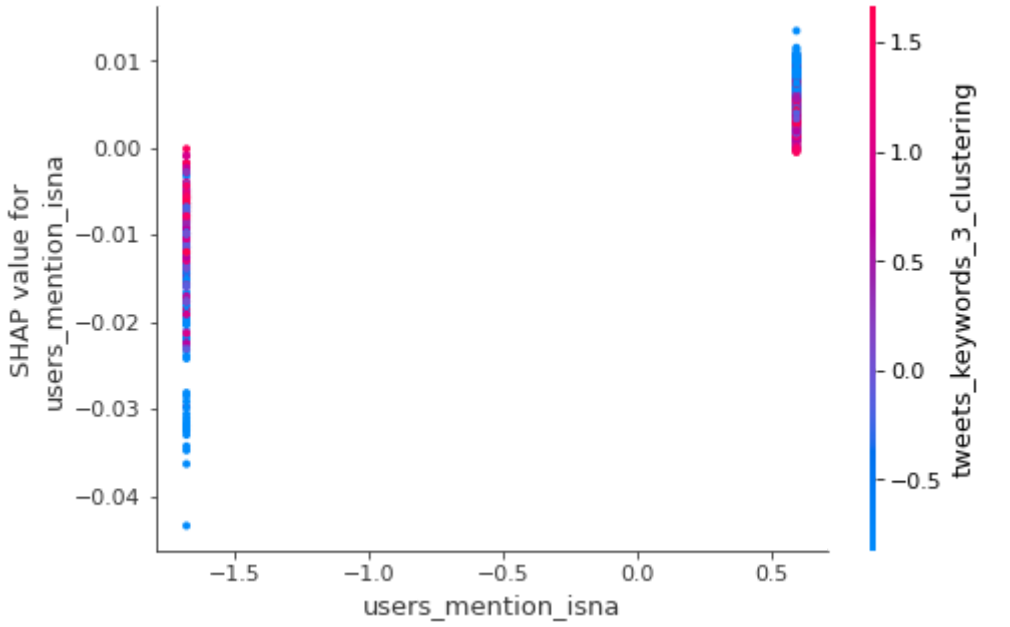

SHAP value for
users_reply_eigenvector_out
0.015
0.010
0.005
0.000
−0.005
0
2
4
6
8
users_reply_eigenvector_out
1.0
0.5
0.0
entities.media

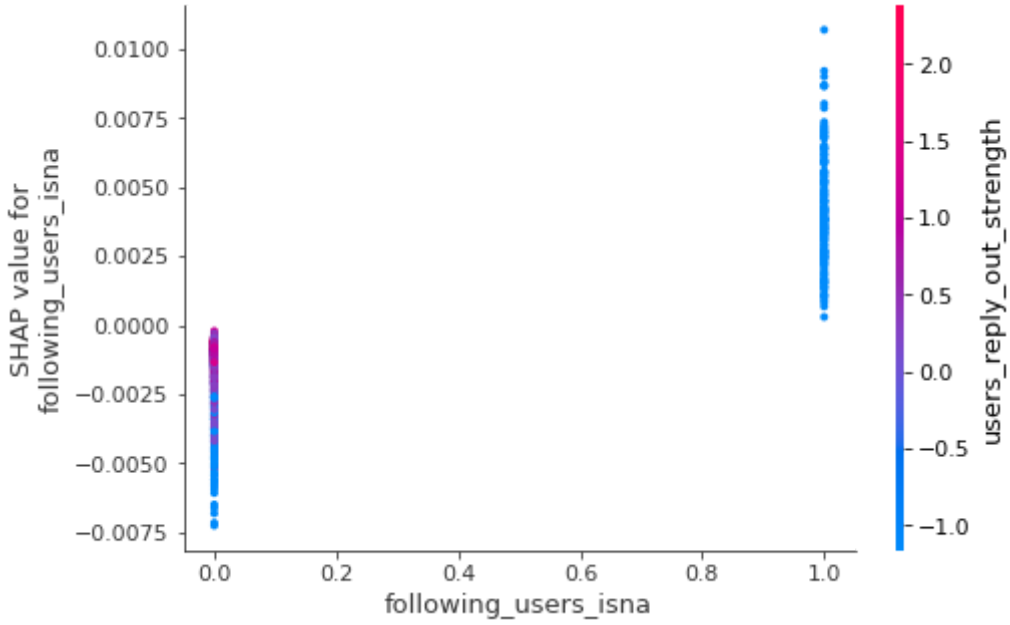

SHAP value for
users_reply_katz_in
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−0.0025
−0.0050
0
2
4
6
8
users_reply_katz_in
1.0
0.5
0.0
entities.urls

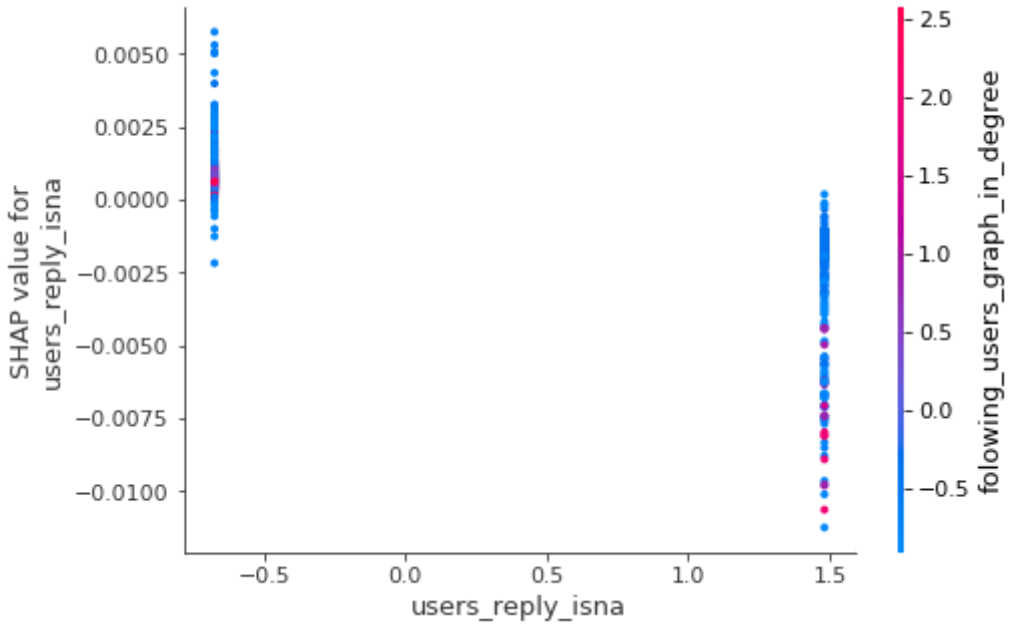

SHAP value for
users_reply_katz_out
0.010
0.005
0.000
−0.005
−0.010
0
2
4
6
8
users_reply_katz_out
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
users_reply_out_strength

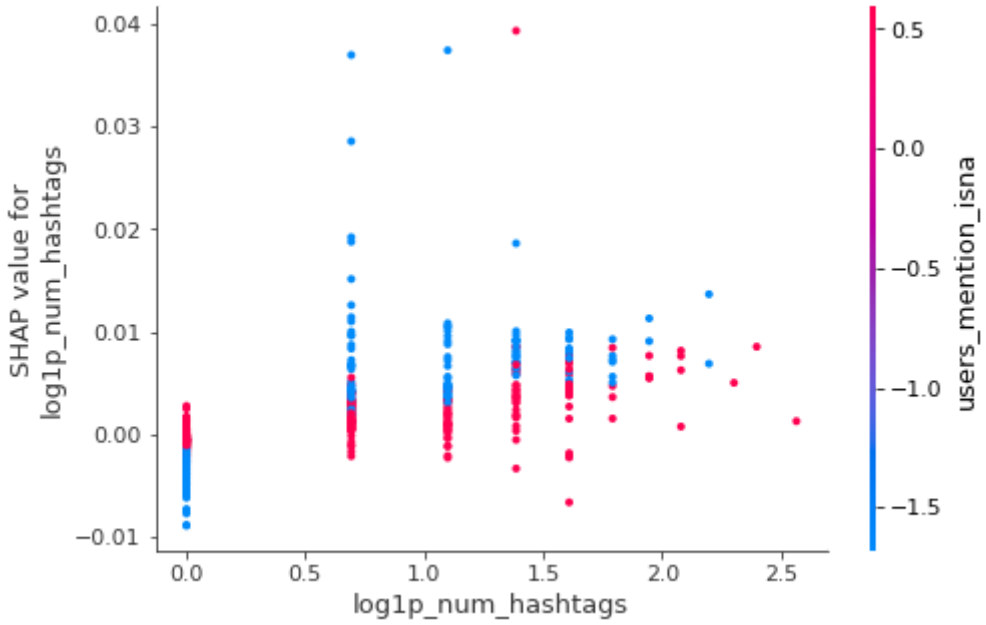

SHAP value for
users_reply_clustering
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.5
0.0
0.5
1.0
1.5
2.0
users_reply_clustering
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
folowing_users_graph_in_strength

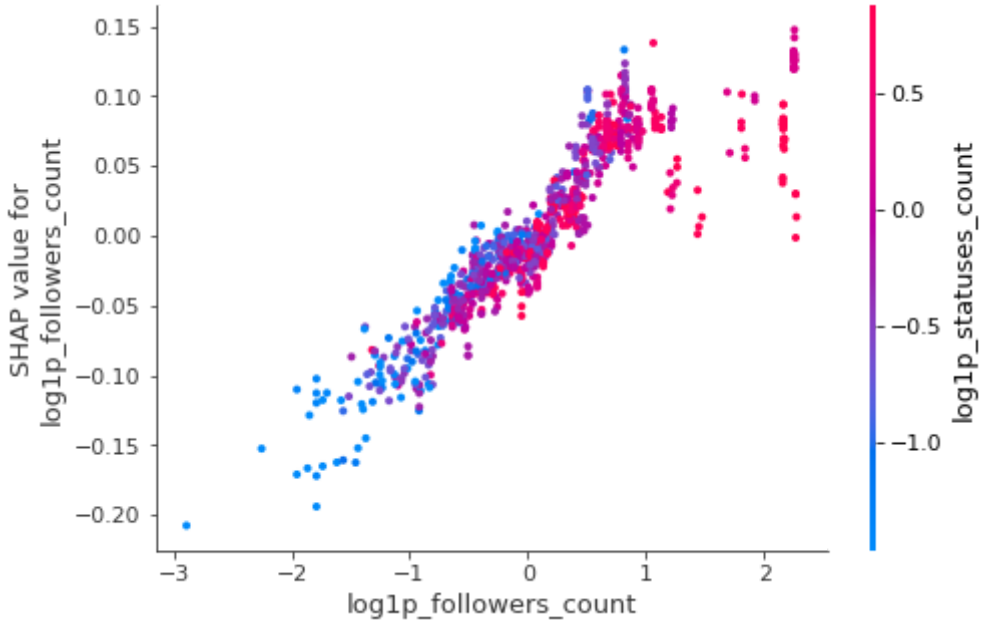

SHAP value for
user.followers_isna
0.0000
−0.0005
−0.0010
−0.0015
−0.0020
−0.0025
−0.04
−0.02
0.00
0.02
0.04
user.followers_isna
0.5
0.0
−0.5
−1.0
−1.5
users_mention_isna

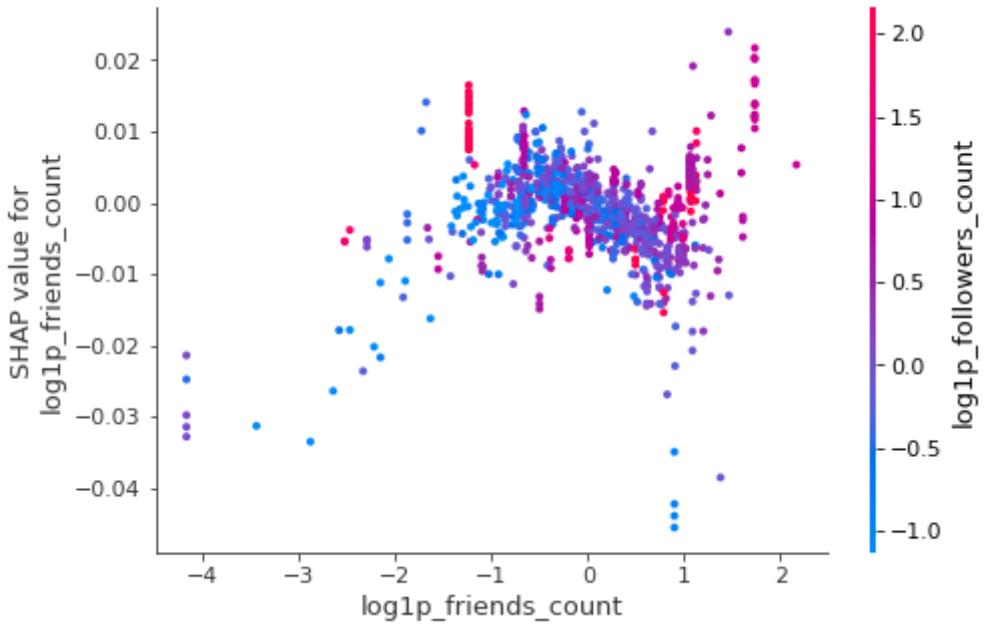

SHAP value for
users_mention_isna
0.01
0.00
−0.01
−0.02
−0.03
−0.04
−1.5
−1.0
−0.5
0.0
0.5
users_mention_isna
tweets_keywords_3_clustering
1.5
1.0
0.5
0.0
−0.5

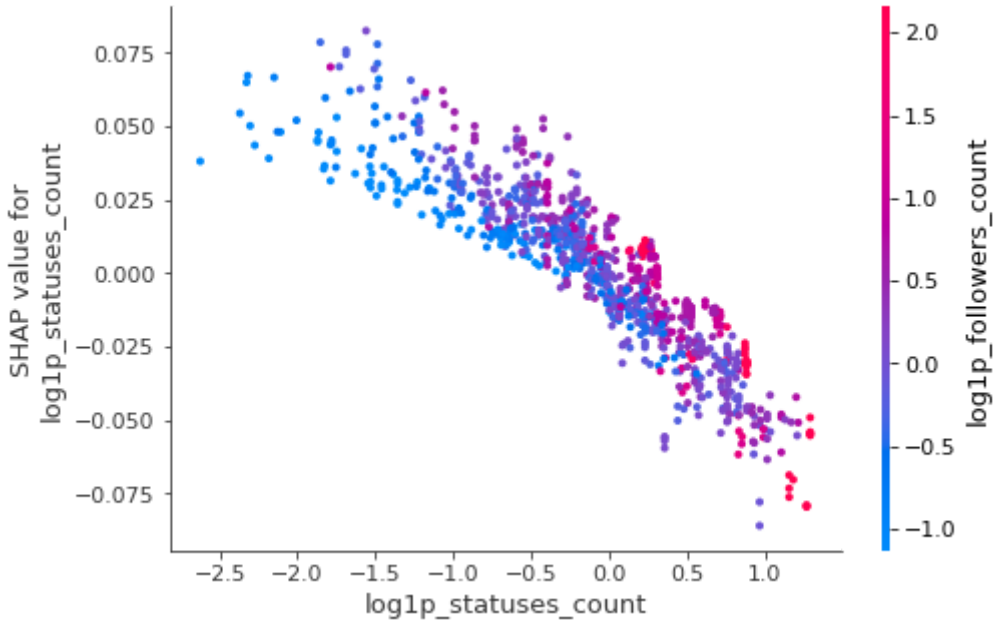

SHAP value for
following_users_isna
0.0100
0.0075
0.0050
0.0025
0.0000
−0.0025
−0.0050
−0.0075
0.0
0.2
0.4
0.6
0.8
1.0
following_users_isna
users_reply_out_strength
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0

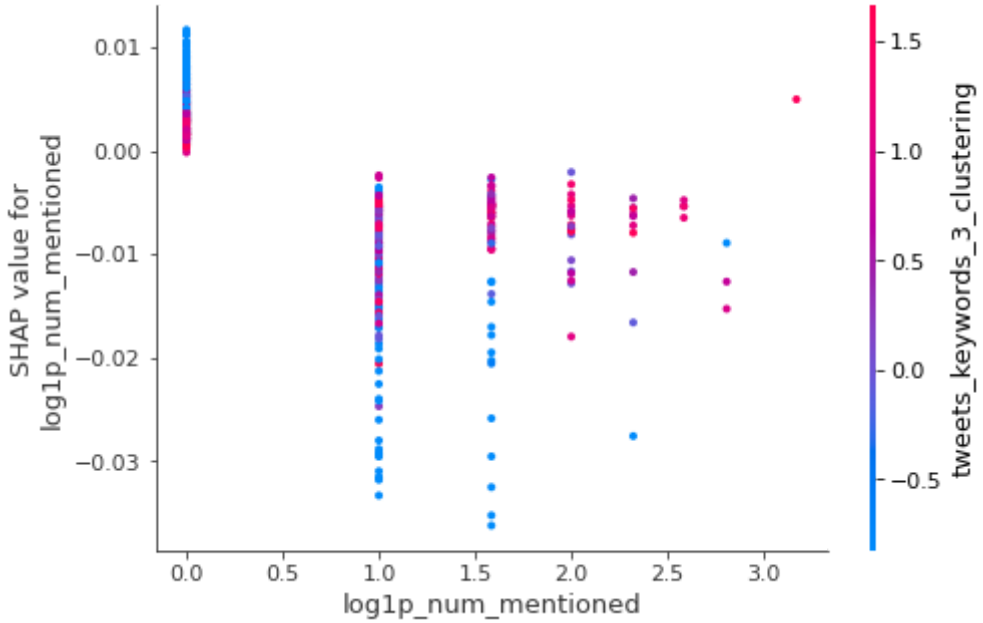

SHAP value for
users_reply_isna
0.0050
0.0025
0.0000
−0.0025
−0.0050
−0.0075
−0.0100
−0.5
0.0
0.5
1.0
1.5
users_reply_isna
folowing_users_graph_in_degree
2.5
2.0
1.5
1.0
0.5
0.0
−0.5


SHAP value for
log1p_num_hashtags
log1p_num_hashtags
users_mention_isna


SHAP value for
log1p_followers_count
log1p_followers_count
log1p_statuses_count


SHAP value for
log1p_friends_count
log1p_friends_count
log1p_followers_count


SHAP value for
log1p_statuses_count
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
log1p_statuses_count
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
log1p_followers_count


SHAP value for
log1p_num_mentioned
0.01
0.00
−0.01
−0.02
−0.03
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log1p_num_mentioned
1.5
1.0
0.5
0.0
−0.5
tweets_keywords_3_clustering