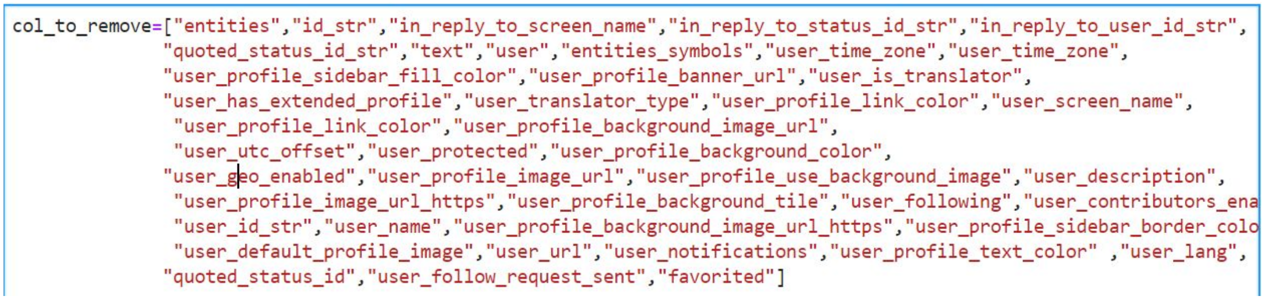
Report on the winning solution in the “Social Media Prediction Challenge”

1) Data Cleaning:

This is where exploratory data analysis was carried in order to understand the data. The documentation provided by twitter was crucial to understand the meaning of each column In the data. This step led to removing many useless features.



```
col_to_remove=["entities","id_str","in_reply_to_screen_name","in_reply_to_status_id_str","in_reply_to_user_id_str",
               "quoted_status_id_str","text","user","entities_symbols","user_time_zone","user_time_zone",
               "user_profile_sidebar_fill_color","user_profile_banner_url","user_is_translator",
               "user_has_extended_profile","user_translator_type","user_profile_link_color","user_screen_name",
                "user_profile_link_color","user_profile_background_image_url",
                "user_utc_offset","user_protected","user_profile_background_color",
               "user_geo_enabled","user_profile_image_url","user_profile_use_background_image","user_description",
                "user_profile_image_url_https","user_profile_background_tile","user_following","user_contributors_ena
                "user_id_str","user_name","user_profile_background_image_url_https","user_profile_sidebar_border_colo
                "user_default_profile_image","user_url","user_notifications","user_profile_text_color" ,"user_lang",
               "quoted_status_id","user_follow_request_sent","favorited"]
```

2) Data processing & features engineering

To improve the performance of this model, additional meaningful features were needed.

List of handcraft features:

- date features (year , month,day of week,hour)
- is_reply_to_status_id: is this tweet a reply to another tweet.
- is_reply_to_user_id: Is this tweet a reply to another user comment on a tweet.
- reply_to_user_id_count: how many times this user received a reply.
- reply_to_status_id_count: how many replies on a status.
- count_entities_user_mentions: the number of users mentioned in a tweet.
- count_entities_hashtags : the number of hashtags mentioned in a tweet
- count_entities_urls: the number of Urls mentioned in a tweet
- len_text : how many words in the text
- Tf-IDF + PCA features : about (20 features) :

`   I used TF-idf in the processed tweets to vectorise them. I also used PCA to reduce the dimensionality of the TF-idf output (This requires a lot of computation resource. Therefore I used a AWS server)

3) Learning phase:

I used Light GBM + XGboost to train the data. I tuned their parameters using grid search. I also used K-fold with (lgbm , xgboost) over 10 fold to slightly improve my score. For the final result, I merged the output of the both models.

4) Provided code:

The solution files are divided into 3 folders along with the model file:

- Data:
    - raw data
    - feautres.ipynp: Notebook for feature engineering
    - text_processing.ipynb: Notebook containing text processing, TF-idf and PCA for tweet text
- data_proc: folder to hold the processed data
- sub: folder to hold the outputs of each model
- Model.ipynp: to train the data