

Dataset Features

We extracted two types of features for the data given on *mediaeval2015/devset/tweets.txt* file, **tweet-based** and **user-based** features, which we stored as comma-separated values (.csv) in **tweet_features.csv** and **user_features.csv** files respectively. The files are associated by the unique *tweet_id* that represents each tweet.

Tweet-based

These features are solely based on the tweet's content. The file has the following format:

```
<tweet_id>, <num_words>, <text_length>, <contains_questmark>, <num_questmark>,  
<contains_exclammark>, <num_exclammark>, <contains_happyemo>, <contains_sademo>,  
<contains_firstorderpron>, <contains_secondorderpron>, <contains_thirdorderpron>,  
<num_uppercasechars>, <num_pos sentiwords>, <num_neg sentiwords>, <num_mentions>,  
<num_hashtags>, <num_URLs>, <num_retweets>
```

For example:

```
263140494127804418, 11, 87, false, 0, true, 1, false, false, true, false, false, 3, 0, 0, 1, 1, 1, 24
```

In more detail, the features are:

- **<tweet_id>** the ID number of the tweet, as given by Twitter.
- **<num_words>** the number of words the tweet text contains, included the hashtags and mentions but excluding the links.
- **<text_length>** the length of characters the tweet text contains.
- **<contains_questmark>** whether the tweet text contains the question mark ('?') character (true or false value).
- **<num_questmark>** the number of the question mark ('?') characters in the tweet text.
- **<contains_exclammark>** whether the tweet text contains the exclamation mark ('!') character (true/false value).
- **<num_exclammark>** the number of the exclamation mark ('!') characters in the tweet text.
- **<contains_happyemo>** whether the tweet text contains any emoticon that claims happy mood (true/false value).
- **<contains_sademo>** whether the tweet text contains any emoticon that claims sad mood (true/false value).
- **<contains_firstorderpron>** whether the tweet text contains the first order pronoun (true/false value).
- **<contains_secondorderpron>** whether the tweet text contains the second order pronoun (true/false value).
- **<contains_thirdorderpron>** whether the tweet text contains the third order pronoun (true/false value).

- **<num_uppercasechars>** the number of the uppercase characters the tweet text contains, excluding the 'RT' acronym that claims that the retweet was retweeted.
- **<num_possentwords>** the number of the positive sentiment words the tweet text contains.
- **<num_negsentwords>** the number of the negative sentiment words the tweet text contains.
- **<num_mentions>** the number of mentions (@username) to other users the tweet text contains.
- **<num_hashtags>** the number of hashtags (#something) the tweet text contains.
- **<num_URLs>** the number of URLs the tweet text contains.
- **<num_retweets>** the number of retweets that the particular tweet was retweeted.

User-based

These features are based on the user's details. Please note:

- If the tweet a retweet of the original one, we refer to the details of the user who retweeted it and not the user of the original tweet.
- The values of the features refer to the ones occurred at the time the tweet was posted. So, the current values of the user's details may be different.

The file has the following format:

```
<tweet_id>, <num_friends>, <num_followers>, <folfriend_ratio>, <times_listed>, <has_url>, <is_verified>, <num_tweets>
```

For example:

```
263101763438120960, 325, 38, 0.11692308, 0, true, false, 1226
```

In more detail, the features are:

- **<tweet_id>** the ID number of the tweet that this user posted, as given by Twitter.
- **<num_friends>** the number of friends of the user (users followed by the user).
- **<num_followers>** the number of followers the user has.
- **<folfriend_ratio>** the ratio friends/followers of the user.
- **<times_listed>** the number that the user was listed in Twitter.
- **<has_url>** whether the user shares a url via his/her profile (true/false value).
- **<is_verified>** whether the user is verified by Twitter (true/false value).
- **<num_tweets>** the number of tweets (posts) the user has posted.