Data Retrieval Process

Data retrieval for this project was from the twitter API which is located at : https:// dev.twitter.com. Twitter has many APIs, each API representing a facet of twitter. They are Twitter for websites, Search API, REST API and Streaming API. The Twitter REST API was used for this specific task.

Initially you are required to set up an account and a password after which you will receive an AccessToken, AccessSecret, ConsumerKey and ConsumerSecret which would then be needed to authenticate you to use the API. You make restful calls using the above credentials.

GET statuses/user_timeline

This returns a collection of tweets posted by the user, who is identified by the user_id or screen_name parameters provided. If the account is protected, you may receive tweets only if your user account has been approved by the user or if you are the user himself.

For example the following call would retrieve the last 200 tweets by the Twitter User named BreakingNews. Many other parameters can be set to retrieve a subset of the data necessary.

"https://api.twitter.com/1.1/statuses/user_timeline.json?screen_name=BreakingNews&include_entities=true&include_rts=true&count=200"

The above call gives us only the tweets by the user, and doesn't necessarily give us a lot of information on the user itself. For this we make a call as follows :

https://api.twitter.com/1.1/users/lookup.json?screen_name=Independent

This would give us the User Information of the user who's screen name is the Independent.

Once data is received this way, we store the last ID for each user in the dataset, such that we then make the next call using the last ID. The data is received in JSON Format, which is then slightly modified to fit our needs and then converted to CSV format to pass on to our Machine Learning algorithm.

We used twenty widely popular News Websites such as the Guardian, BBC, CNN etc and found around twenty five fake accounts to retrieve untrustworthy data. We have a dataset with about 40,000 tweets, collected over 5 days. The above code was written in java and Manipulation of the received raw data was done using the north concepts library (which is explained in section x).

JSON Results :

The raw JSON that we receive from the REST calls have plenty of information that we may not need, some of them have a significant amount of null values which might cause problems in our analysis later. For this reason some features of the returned results were dropped at various stages in the program. (Some at initial data collection level and some later on in the actual dataset used for measure). The following features were considered useful / important for the different aspects of this program.

Features Received :
Features deemed important : CreatedAt, ID, ActualTweet,Source,Truncated,InReplyToStatusID

Features used for Logistic Regression : CreatedAt, InReplyToStatusID, InReplyToUserID, UserID, RetweetCount, FavouriteCount, Hashtags, URL, MediaURL, MediaType, UserMentionID, PossiblySensitive, TweetLength.

Features used for SVM :

The following table explains the features in detail and the reasoning behind the inclusion or exclusion of this information in our classifiers.

Tweet Features :

Feature Name	Explanation	Included In	Reason
Created_At	UTC time for when this tweet was created	LR, SVM	This could be used to see if frequency has an impact on trustworthiness. For instance if the number of tweets per minute has an impact on the trustworthiness of a user
Id	Tweet ID	-	The ID of the tweet could in no way impact our analysis
Id_str	Tweet ID in String Form		As above
Text	Actual Tweet	-	Unless analysing the contents of the tweet this information is not important. However we have a derived feature which is the length of the text string which is used in the analysis.
Source	This is the Utility used to post the tweet, as an HTML-formatted string.	-	Although it will be interesting to see if there is any correlation between the utility used to post the tweet with trustworthiness of the tweet, most tweets are posted from the twitter web client which has a source value of 'web' hence this wouldn't make a huge difference.
truncated	If the tweet was truncated or if its the original length	-	The Twitter API says that this feature is rarely ever used since twitter now rejects long tweets as opposed to truncate them - we saw no reason to use this.
in_reply_to_Status_id	If this tweet was a response to a status then the integer representation of the ID of the status	LR, SVM	This would tell us if its a reply or a tweet.
in_reply_to_Status_id_str	If this tweet was a response to a status then the string representation of the ID of the status	-	This is just a repetition of the previous feature in String Form
in_reply_to_user_id	If this tweet was a response / message to a user, the integer representation of the UserID	LR, SVM	This feature can be used to check to see if it is a response to a user or a tweet.

Feature Name	Explanation	Included In	Reason
in_reply_to_user_id_str	If this tweet was a response / message to a user, the string representation of the UserID	-	Repetition of previous information above.
in_reply_to_screen_name	If the tweet is a response then this field will have the screen name of the original author	-	Repetition of previous information above.
user	This is an embedded JSON array containing information about the user who posted the tweet.	-	Since we made another call with the actual user ID to retrieve information about the user this array will just be a replica of what we already have. The advice from Twtitter API is also such that embedded information such as this are unreliable.
geo	Deprecated	-	Although this appears as part of the JSON received this feature is deprecated in this API.
coordinates	The geographical location of the tweet if provided.	-	In almost all the data we received this information is not present hence we decided to drop it. However user locations are available on the user profile.
place	If value is present then this shows that the tweet is associated but not necessarily originating from this place.	-	Again a huge majority of the samples we received didn't have this information.
contributors	A collection of brief user objects showing the people who contributed to the authorship of the tweet.	-	This will only work if the user has turned the contributors enabled feature on. This is a beta feature and is not available to all the users at this point of time, hence i decided to not include this in our analysis.
retweet_count	Number of times a story has been retweeted	LR, SVM	The number of times tweets are generally retweeted can be used to check if there are differences in our data.
favourite_count	Number of times a story has been favourited	LR, SVM	The number of times tweets are generally favourited can be used to check if there are differences in our data.
entities	Entities provide information about the text of the tweet without having to parse the text to find out the structure. It shows you URLs, MediaURLs, Hashtags and symbols used in a tweet.	LR, SVM	This information is deemed valuable as we can analyse the differences in structure of a trustworthy tweet with an untrustworthy one.
favourited	This indicates if the tweet has been favourited by me (Authenticating User)	-	This piece of information is irrelevant, since it concerns the authenticating user (me in this instance) hence the exclusion.

Feature Name	Explanation	Included In	Reason
`retweeted`	This indicates if the tweet has been retweeted by the authenticating user.	-	This piece of information is irrelevant, since it concerns the authenticating user (me in this instance) hence the exclusion.
`possibly_sensitive`	This field only comes up if the tweet contains a link. This is an indicator, that the URL in the tweet may contain media that may be sensitive.	LR, SVM	A value in this field indicates two things : That a URL is present and if the URL is sensitive or not. Hence this feature could be useful for our analysis.
`lang`	This indicates a BCP 47[1] language identifier which shows the language that was detected by the machine for the tweet.	LR, SVM	This feature is used - since we want only tweets that are in english.
`frequency`	DERIVED : This is a derived value from. We Take the first and Last tweets of a User Over a time period and calculate the frequency of the tweets	LR,SVM	This could be an important indicator in informing us patterns for different types of users, hence we use this feature.
`IsItARetweet`	DERIVED : This is a column that checks if the tweet is a retweet or an original.	LR,SVM	This could be used to weed out duplicates.

User Features :

Feature Name	Explanation	Included In	Reason
id	Unique Identifier for User	-	Not very useful for analysis
id_str	String representation of the id above	-	Not very useful for analysis
name	The name of the user as they've defined it	-	Not very useful for analysis
screen_name	The name a user is identified with on twitter	-	Not very useful for analysis
location	The user defined location for this account.	-	Including this diluted our analysis since we had profiles from all over the world, and we only had 40 profiles in total.
description	A string describing the users account	LR, SVM	The description itself is not used in our analysis however the length of the description string is used in our analysis.

Feature Name	Explanation	Included In	Reason
url	A URL associated with the user provided by him/her	LR, SVM	We check if the user has a URL or not. We don't use the URL itself.
entities	Entities that have been parsed from the URL or description fields.	-	This information has already been taken into account from other features.
protected	This indicates if the user has chosen to protect their tweets	LR, SVM	This might have some significance as well hence included.
followers_count	The Number of followers this User / Account currently has	LR, SVM	This is a very important feature, hence it is included in both our classifiers.
friends_count	The Number of users this user follows on twitter	LR, SVM	This is also a very important feature hence it is included.
listed_count	This is the number of public lists this account is a member of.	LR, SVM	This is also a very important feature hence it is included.
created_at	The date and time this account was created	LR, SVM	This feature also might be an indicator of the trustworthiness of the source/ account.
favourites_count	The amount of tweets this user has favourited from the date of opening the account	LR, SVM	It will be interesting to see if there are any patterns in how active the user is in terms of favouriting other tweets and responding to trustworthiness.
utc_offset	This is the offset from GMT given in seconds.	-	This is not used since Location has already been used once.
time_zone	This string gives the time zone within which the user profile is within	-	Once again this feature would be a duplicate of Location hence we did not use it.
geo_enabled	If true this means that the user has enabled geo tagging for his tweets.	LR, SVM	Again, it would be interesting to identify any patterns amongst the different types of users.
verified	Verified twitter accounts are accounts which twitter has verified - stating that the user is who they say they are on twitter.	LR, SVM	This could be a very important factor in predicting trustworthiness hence it is included in our model.
statuses_count	The number of tweets published by the user.	LR, SVM	This again could prove important - we just need to check for patterns.
lang	The BCP 47[1] code for the language used by the user. This is self declared.	-	This feature is not used since it could negatively impact the analysis, or create a bias that is unnecessary.
status	This gives information about the users most recent tweet.	-	We will not be using this feature since we have all the necessary information we need about the users tweets from the other call.

Feature Name	Explanation	Included In	Reason
contributors_enabled	This allows the tweets by this account to be coauthored by another account	-	According to Twitter, this is rarely true. Hence this feature is ignored.
is_translator	If this is true, then it indicates that the user is a part of the translator community for twitter	-	This feature again is rarely true hence this wasn't included in our model.
is_translation_enabled			
profile_background_color	This is the hexadecimal color chosen by the user for their background.	-	We didn't think that the choice of colour could be impactful since each account will use colours they are recognised with.
profile_background_image_url	A HTTP-based URL which directs you to the URL of the image that the user used on his profile	LR, SVM	This feature was used, although not as the URL. We modified this attribute to boolean, true and false being profile background image exists, and profile background image doesn't exist respectively.
profile_background_image_url_https	This is the same information as the row above	-	Repetition
profile_background_tile	This indicates that the users background image should be tiled	-	This is an attribute that concerns the aesthetics of the profile hence not considered.
profile_image_url	A HTTP based URL directing you to the users profile image	-	Most profiles, have an avatar - if they tweet significantly. Hence this information was not considered.
profile_image_url_https	A HTTP based URL directing you to the users profile image	-	Repetition
profile_link_color	The colour that the user chose to display links.	-	Again, this concerns the aesthetics of the profile, hence not considered.
profile_sidebar_border_color	The colour that the user chose to display the borders of their profile	-	Again, this concerns the aesthetics of the profile, hence not considered.
profile_sidebar_fill_color	The sidebar fill colour	-	Again, this concerns the aesthetics of the profile, hence not considered.
profile_text_color	The colour for their text	-	Again, this concerns the aesthetics of the profile, hence not considered.
profile_use_background_image	If true the user wants to use a background image.	LR, SVM	It would be useful to see if there are any patterns.

Feature Name	Explanation	Included In	Reason
default_profile	This if true indicates that the user has not changed the theme or background of their profile.	LR, SVM	It would be useful to see if there are any patterns.
default_profile_image	If true this indicates that the user has not changed the avatar for his profile	-	This is loosely covered in some of the features above.
following	Shows if my profile is following this user	-	This piece of information is not significant.
follow_request_sent	Shows if I have sent a request to follow this user	-	This piece of information is not significant.
notifications	Indicates if the user has chosen to receive his tweets via SMS	-	This is deprecated hence not used.

Content Based Features : TweetLength, NumberOfSmilies, TweetEmotion, BREAKING, SHOCKING

Data Visualisation

TODO : Write about the visualisation code if you have any.

Visualise in black and white being 1 or 0 the Sad, Happy count. Also emoticon.

http://www.washington.edu/news/2014/03/17/hold-that-rt-much-misinformation-tweeted-after-2013-boston-marathon-bombing/

http://www.geekwire.com/2014/boston-marathon-twitter-error-researchers/