The Proposed Model

This part illustrates the design description for the proposed arabic rumors detection technique (classifier).

The rumors tweets will be defined as false information for social events that generates many social responses by using re-tweet and replay options.

1.Functionality and Design

The process of designing detection rumors classifier can be divided into five basic categories.

They are as follows:

1. Data Acquisition.
2. Data Preprocessing.
3. Data Labeling.
4. Feature Extraction.
5. Classification.

1.1 Data Acquisition (Collection)

The tweets were collected using Twitter streaming API to collect data in real-time from January 1 until April 1, 2018. The tweets were collected by using different track values for filter stream function. The values of the track are:

- Trending hashtags in Saudi Arabia from different domains (political, social, sports, technology, and economic).
- Using Specific keywords to track tweets like persons' names, cities or events.

1.2 Data Preprocessing

After the tweets were collected and saved, the preprocessing step was started by:

- Removing duplicate tweets.
- Deleting some values that we did not need such as: tweet id, user id, profile image.
- Converting some values to Boolean values such as
    - Location and description based on if they have value or not.
- Performing some calculations include:
    - Account age= current date - the account creation date.
    - Number of Statues/account age
    - Number of followers/number of friends
    - Number of favorites/number of statues
    - Time Span= the tweet creation date – the account creation date.
    - Engagement Score= number of statues/Time span.

1.3 Data Labeling (Annotation Process)

There were two copies of tweets collection that were annotated (labeled) by two individuals to either rumor or not-rumor. For each part of data collection, the links of tweets were sent to the annotators to help them to display tweets on twitter and classify it to either rumor or not rumor. The annotation process will be based on:

- Comparing the tweets' contents with credible sources of the news like online newspapers, official twitter accounts of ministry and official spokesperson of a ministry or a company.
- Determining the relation between tweets' contents and the recent events that are happening by searching for the newest news and events.
- Looking for trustful sources that confirm the tweet that was propagated or asking them about the truthfulness of the tweets.

- Searching for the tweets contents to find out if the contents were mentioned last year or two years ago and reading the articles and information that relate to the contents.

After the annotators completed their annotation, two techniques were used to measure the agreement degree among them.

1. Each annotator had 5% of repeated tweets in his part of the collection, which will be discarded later from the data set to avoid any confusion on the classifier execution side.
2. Using the formula below

Formula: agree (K) =(Pr(a)- Pr(e)) / 1- Pr(e) , where Pr(a) is relative agreement and Pr(e) is probability agreement chance .

An extra technique was used to increase the level of the agreement, by asking an extra annotator to make final judgment on the tweets that two annotators assign different labels for them.

1.4 Feature Extraction.

The following features were extracted from a user, account, and tweet's content which will be used to train the model of rumor tweets detection. The features were chosen based on some previous studies on detecting rumors' tweets and empirical analysis that was performed on a sample set of Saudi rumors' tweets that collected from trending hashtags.

- User Based Features
    - User Name: a name of the user that was written in the account profile (personal, organization).
- Account Based Features
    - Verified: account is authentic, it will have a blue badge appear next to the name on the account's profile. It has two values either True or False.

- Account protection: it has two values: True or False.
- Account Age: the time passed since the user created the account.
- Description: some information written in the account's bio.
- Location: indicates the location of the user that entered in his/her profile.
- Number of Followers: the number of people who follow and read the user's tweets.
- Numbers of Friends: the number of people who the user is following.
- Number of Statuses: number of tweets that the user posted.
- Favorites Number: the number of tweets the user liked.
- Number of Lists: the number of lists that a user belongs to.
- Account Name: the name that is written after @.

- Content Based Features
  - Question Mark: checking the content if it is included ? or not.
  - URL: the tweet includes the URL and the numbers of it.
  - Multimedia: the tweet includes images, videos and the numbers of them .
  - Hashtag: the tweet includes hashtags and numbers of it.
  - Replay/post/re-tweet: to know if the tweet is replied tweet or posted tweet by user or is re-tweeted by the user.
  - Number of re-tweets: the number of re-tweets for the tweet.
  - Number of likes: the number of likes for the tweet.

- Engagement Score: the number of tweets divided by the number of days since the user account creation.
- Time Span: the difference between the date of tweet and the account registration date.
- Phone Number: a tweet includes a phone number that started with (05).

In addition to all the features above, there are some calculations that must be computed which will help in deciding the tweet's class. They are:

- Number of Statues / account age
- Number of followers / number of friends
- Number of favorites / number of statues

A. <u>The reason for choosing some features in detecting rumor tweet:</u>

- Protected and verified accounts do not post or re-tweet rumors.
- Rumors tweets did not include phone numbers , which can be helped to distinguish between rumor and spam tweets.
- Rumors tweets did not include question marks, the question mark can be found in the replies to the rumor tweet.
- The number of followers, friends, and lists can be used as indicators to detect rumors.
- When the name of the user is real and the account name is a name of organization or matching the real name of the account's user this will reduce the probability to propagate or post a rumor.
- Rumors tweets include URL, multimedia and hashtags.

- The number of days since the account creation can be compared with the number of statuses that the user posted.

1.5 Classifier Training

After the whole processes above were completed, the classifier was trained by using many machine learning algorithms that can be used to learn the model.