Multinational Mood Detection Based on Tweets in English and Spanish

Adnan Vilic
Technical University of Denmark
+45 – 31 55 88 22
s052240@student.dtu.dk

Alina Ioana Manolache
Technical University of Denmark
+45 – 50 28 13 49
s101373@student.dtu.dk

ABSTRACT
This paper describes the final project from the course 02817 – Personalization and Metadata Models, at Technical University of Denmark. The project investigates a method for detecting the mood each country has towards the rest of the world. This is done by extracting tweets that mention at least a country's name and then analyzing the tweets using sentiment analysis in Spanish and English. Additionally, image analysis is used to visualize the results on a world map.

Categories and Subject Descriptors
I.2.7 [**Natural language processing**]: Text Analysis

D.m [**Miscellaneous**]: Data mining

General Terms
Design, Experimentation, Human Factors, Languages, Theory.

Keywords
Natural language processing, affective norm sets for words (ANEW), Twitter, MATLAB, .NET, data mining, image analysis.

Contributions
The project has been done together in the team. In the report Alina Ioana Manolache wrote sections 1, 3 and 5 while Adnan Vilic wrote sections 2 and 4.

1. INTRODUCTION
By applying sentiment analysis, large amounts of data can be processed to obtain an overview of the mood in the text. Sentiment analysis has many applications, such as understanding the emotional development of a story, reputation of a company in the news, or opinions about candidates during elections.

In this project, sentiment analysis is used to track the mood around the world by extracting tweets and analyzing those using affective norms for words (ANEW) in English and Spanish. For each country, tweets are extracted for what its residents tweet about the rest of the world.

To visualize the results, image analysis methods are used to display the moods on a world map.

2. DATA MINING
Currently, there are no existing datasets with Twitter sentiment messages or tweets with metadata about where the tweet was posted from.

To obtain usable datasets, a scraper was developed in .NET which downloads tweets based on several arguments:

- Language in which the tweet was posted.
- Coordinate as latitude and longitude of posted tweet.
- Radius in kilometers around coordinate from where the tweets are gathered.
- Country for which the tweets are extracted.

For each country, a geographical coordinate represented as latitude and longitude and, along with a radius is used to identify from which country a tweet was posted. Figure 1 shows an example on locations from which tweets are extracted. It shows areas that identify Iceland, Denmark and France. As seen from circle around France, it is not always possible to cover the entire country without overlapping a different country. Therefore, the largest possible circle is created for each country including major cities in the country.

Figure 1. Map of Europe. Circles are examples on areas from which tweets are extracted

Once all locations and radii are specified for each country, tweets can be downloaded. For each coordinate with a radius, Twitter search engine is used to download all tweets that refer to other countries around the globe.

Using this approach, two datasets have been generated for tweets in English and Spanish for the largest 166 countries in the world. The datasets consists of:

- 270.000 tweets in English
- 78.150 tweets in Spanish

The English dataset contains only tweets for the period 25th April to 30th April 2011, while the dataset for Spanish tweets is for the period 26th April to 2nd May 2011.

3. SENTIMENT ANALYSIS
To evaluate the mood of each tweet in the dataset, affective norms sets for words in English [1] and Spanish [2] are used.

After evaluating all tweets by applying the two ANEW sets, the total amount of tweets that can be used for sentiment analysis are:

- 124.300 tweets in English
- 18.000 tweets in Spanish

The Spanish dataset is greatly reduced because a large amount of words that usually have characters like á, í and ó, are sometimes spelled without an accent in the tweets.

For each country in the world, the average valences are calculated which represents the country's mood towards the rest of the world.

3.1 Valence Distribution and Average

Statistical methods can now be applied to understand the content of the gathered datasets.

Figure 2 shows the overall valence distribution for all countries in Spanish and English. The distribution shows that the majority of tweets for countries have high valences in English and only a few countries are below or on the average. For tweets in Spanish, countries still have high valences, although closer to the average. The reason for this is partly that a part of the Spanish dataset was generated after Osama bin Laden's death was announced on 2st of May 2011.

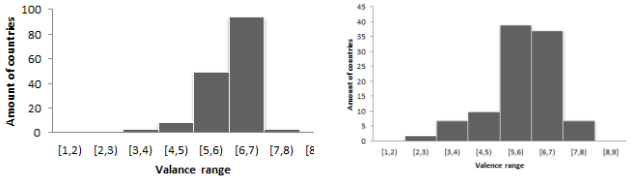

Figure 2. Valence distribution for tweets from all countries in English (left) and Spanish (right)

It is also possible to see which countries have the overall most positive or negative tweets towards other countries.

The average mood that a country, *x*, has towards other countries is calculated as:

$$x_{avg} = \frac{\sum avg(c) * f(c)}{N}$$

where avg(c) is the average valence towards a country, c, and f(c) is the amount of tweets towards the country. N is expressed as the sum of all tweets posted from *x*.

MOST POSITIVE MOOD TOWARDS OTHER COUNTRIES				
English	Iceland (7.21)	Mali (6.95)	Honduras (6.91)	Haiti (6.89)
Spanish	Honduras (8,41)	Puerto Rico (8.36)	Belarus (7.29)	Venezuela (7.06)

MOST NEGATIVE MOOD TOWARDS OTHER COUNTRIES				
English	Iraq (4.51)	Pakistan (4.73)	Ecuador (4.86)	Qatar (4.91)
Spanish	New Zealand (3.89)	Syria (4.48)	Portugal (4.81)	Greece (4.87)

Figure 3. Most positive and negative moods towards other countries. Average weighted valence listed below country.

The average valence for all Spanish tweets is 6.04, while the average valence for the tweets in English is 5.99. These values are very close to one another and in the upper range of the valence scale. This is due to the fact there are more words in ANEW sets with a positive valence than with a negative one.

3.2 Sufficient Data

One of the important factors when determining the mood of the countries is the amount of data available. If only one tweet is written about a country, one should not take it as the representation of the mood for the whole country.

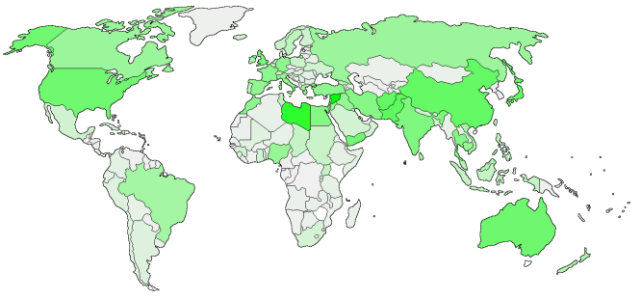

Figure 4. Amount of tweets in English posted about each country

The saturation of the green color in on the world map in Figure 4 shows how many tweets were posted about each country. The higher the saturation, the more tweets were posted about the country.

The most talked about countries on Twitter in the investigated period are Syria, Libya, Pakistan, Afghanistan, China and Japan. All these countries have had important political or social events during April 2011.This leads to the conclusion that once a newsworthy event occurs in a certain country, that the particular country captures the attention from people on Twitter as well.

4. VISUALIZATION

For a better overview, the valences for each country can also be visualized on a world map. For this purpose a world map with political borders has been obtained from [3] (see Figure 5).

Figure 5. Empty world map with with political borders

To plot the valences for each country in the map, the map needs to be preprocessed using several image analysis algorithms. This is done in MATLAB.

4.1 Preprocessing the Map

First, the image is binarized to make all pixels in the image either white or black and to remove all different shades of the two colors.

Then connected component labeling [4] is applied to find the areas representing each country. The algorithm identifies and labels all pixels within an object with the same color. It compares the color of each pixel with every of its neighbors. To determine whether two pixels belong to the same object 4-connectivity is

applied as shown in Figure 6. The blue color represents the pixel currently investigated. If the orange pixels (neighbors) have the same color as the blue, they belong to the same component.

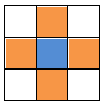

Figure 6. 4-connectivity

Once the labeling is done, each connected component, representing a country or water, will have a unique number. For each country, a pixel coordinate in the image was manually selected, which represents a point within the area of the country.

Now the country can be colored by extracting the value of the pixel within the area (connected component) and replacing all pixels with that value with the desired valence color.

4.2 Determining Valence Color

Before valence can be plotted inside the image, it needs to be converted to a color. The valence color is in a color gradient between green and red:

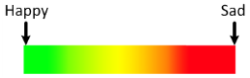

Figure 7. Valence gradient

Given a valence value, the valence color is determined as:

```
sad   = [255, 0, 0]
happy = [0, 255, 0]
r = sad(1) + (x * (happy(1) - sad(1)) / 9)
g = sad(2) + (x * (happy(2) - sad(2)) / 9)
b = 0
Valence color = (r,g,b)
```

where r, g and b are respectively the red, green and blue saturations of the new valence color, sad and happy are the colors for happiest and saddest valences, 9 is the amount of different valence levels, and finally x the calculated valence for the country.

4.3 Plotting Valences

After preprocessing the world map and assigning a color for each valence, the valences can be plotted for all world countries. The valence for one country is based on the average mood of the rest of the world about that specific country. This means that if a country is colored in green, people from the rest of the countries in the world have tweeted positively about it, as in Figure 8.

The valence of a country can also be based on the average mood coming from only several specific countries. For example, it is possible to visualize the mood that countries in Europe have towards the rest of the world.

Another option would be to visualize what a certain country tweets about other countries around the globe. In this case, a country being green means, that the country, that is taken as point of reference has mostly positive tweets about it.

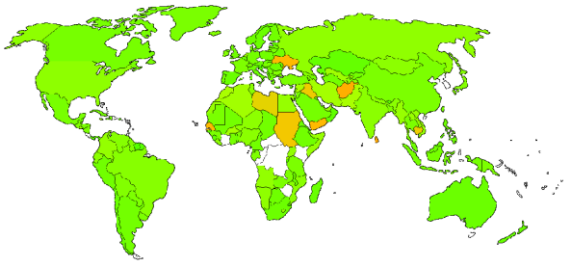

Figure 8. World mood for tweets in English

Figure 8 shows the general world mood based on English tweets for all countries. Several countries such as Czech Republic, South Korea, North Korea and several African countries are colored white, as nothing was tweeted about these countries using same terminology as in the developed scraper.

For each of the yellow and orange countries, there are clear historical events that caused the valences:

- **Afghanistan, Iraq and Libya**: Currently in war.
- **Sudan:** Conflicts got worse close before elections.
- **Sri Lanka:** A report emerged on genocides committed against Tamils on 29th April 2011.
- **Thailand:** Massive anti-government demonstration on 30th April 2011.
- **Ukraine:** 25th anniversary since the Chernobyl accident on 26th April 2011.
- **Yemen**: Protests against the government.

From the global valences of tweets, most having high values, we can deduce that people mostly tweet positively about other countries, unless a bad event happens there.

Figure 9 and Figure 10 show valences of tweets from USA and Iran in respect to the rest of the world.

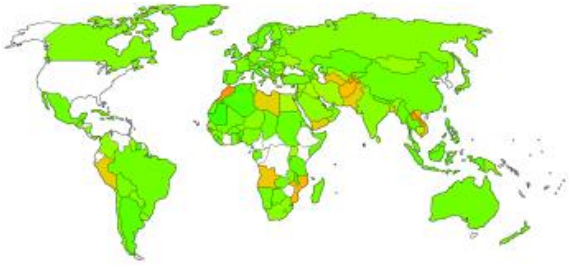

Figure 9. USA's mood towards rest of the world

USA's mood towards other countries (see Figure 9) shows that Iraq is green even though USA is in war there. This is due to the Americans in general supporting the war and their troops in the area. While there are supporting tweets for Afghanistan as well, it still gets a negative valence due to death of an American soldier during the period in which tweets were gathered.

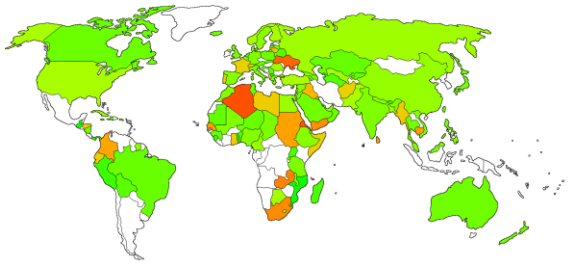

Figure 10. Iran's mood towards rest of the world

Looking at Iran's mood towards other countries (Figure 10), we see that Iran has a positive attitude towards USA despite having many political disagreements with them. In this case it is very important to be cautious about the interpretation of the results. This is because sentiment analysis is conducted on English, and only English speaking, tweeting Iranians are taken into account. If sentiment analysis had been performed in Farsi, which is the national language of Iran, the view would also take people into account who are against USA.

From the Iran case it can be seen that in order to be able to estimate the mood of a population one must have a reasonable representation within the samples taken. Considering that most anti English text and propaganda would be aimed at the majority of the population it would be written in Farsi and thus not taken into account when performing the sentiment analysis.

Another interesting point is that the people that are normally using Twitter in a country like Iran would have to be wealthier that the average population which in most cases also means more educated. The level of education normally also has an effect on peoples mood and peoples attitude towards other cultures and nationalities.

This can be compared with for example Denmark, where everyone has access to a computer and the education and wealth would not have influence on the access to the Internet. However again the language is a problem since the majority of the population speaks their mother tongue.

The solution to the above problem would be to create multiple affective norms for words that contain valences for words in the three or four most used languages for each country that is analyzed. Furthermore a weighting and normalizing between the amounts of tweets submitted in each language would be necessary in order to have a more correct picture of the mood in the country. Finally normalization using statistics and estimates of the internet users within population could be used for having some kind a un-certainty estimate of the determined mood.

5. CONCLUSIONS

In this project, a rough model has been developed for tracking mood multinational moods based on tweets in English and Spanish. While it can still be improved by expanding the area for each country from which tweets are extracted, as well as through applying ANEW for other languages, the application still allows tracking the most important news at a given time. As people tend to tweet in their mother tongue, applying sentiment analysis for tweets in all or at least the most spoken languages in the world (Mandarin, Hindi, Arabic, Russian, German etc.), would have been a big plus for this project. Unfortunately it was possible to obtain affective norms for words in other languages, except English and Spanish.

As ANEW consists of slightly positive words, the tracking mainly works for negative news. As an example, positive news such as the British royal wedding on the 29th of April is unnoticed in the visualization, with UK being just as green as majority of the countries in the world. On the other hand, the conflicts in Libya stand out.

It would be interesting to be able to choose the period for which sentiment analysis should be applied, in order to cover specific events such as the revolution in Egypt from 2011, or FIFA World Cup 2010. Unfortunately, Twitter's search engine does not allow searching for specific periods.

Additionally, the emotional development in tweets over time could be tracked, to see how accurately sentiment analysis can be used to detect news with biggest impact, such as the time when the earthquake hit Japan in March 2011, or see how voters' opinions regarding the candidates changed over a prudential campaign.

5.1 Perspectives

The project has shown some potential for being used in the future in order to estimate the mood of the people in different countries and as such it could be used to predict results of events where people's opinion has important role. Examples could be who is going to be winner of Eurovision Song Contest, presidential elections or to determine reputation of companies in order to use it for prediction of stock market development.

6. REFERENCES

[1] Margaret M. Bradley, Peter J. Lang, 1999, *Affective norms for English words (ANEW)*, http://www.uvm.edu/~pdodds/files/papers/others/1999/bradley1999a.pdf

[2] Jaime Redondo, Isabel Fraga, Isabel Padrón, *2007, The Spanish adaption of ANEW (Affective Norms for English words).* http://www.springerlink.com/content/c70114l45436530w

[3] *World map*. http://outline-world-map.com

[4] R. Fisher, S. Perkins, A. Walker, E. Wolfart, 2003, Connected Component Labeling: http://homepages.inf.ed.ac.uk/rbf/HIPR2/label.htm