Project Proposal

Different countries…Different people…Different opinions

Sarat Kiran Andhavarapu
Shanmukh Yadlapalli

Introduction

The focus of the proposed project is to investigate how people from different countries react to a situation/event and also to determine if the opinions of people change with time and/or within countries for a similar situation/event. With advancement in the field of communication, people across the globe are able to get the same information and express their feelings, at the very moment as an event occurred in any corner of the world. Twitter is one of the most popular platform for finding real time information such as news, people's reaction and discussions. In the proposed project we analyze twitter data-set to achieve our goals.

The first step toward achieving the proposed research is to collect all the tweets people posted to a present situation and similarly, tweets that are made when the same kind of situation that occurred in the past. Here, we are accounting the tweets only when the tweet contains location information. People using smartphones to tweet can provide much precise information about their location.

Here, we take twitter data-set related to newly released IPHONE 5S. We intend to collect the tweets of people who have expressed their feelings about the phone for a certain time period. To ascertain that we include everything to analyze peoples feeling, we take following things into account -emoticons, terms they used in their tweet, geo-location, Hashtags, and pictures they attached/used with their tweet. We aim to keep all the people who tweeted in three categories, who feel good, bad and neutral about the iphone and then to sub categorize them according to their location information.

Previous research work ?? what remains unknown ??

We have looked into various papers (placed in the reference section) and did not find a paper that uses all the features mentioned in our proposed project. However, we found some methods to evaluate each feature separately in the mentioned papers, we are going to use the methods (aim to enhance them) for evaluating some of our features.

Why this is interesting ?? why do we care ??

In the future work, we can even determine the reasons for the change of opinion. Lets suppose, If major people in INDIA felt the iphone 4S is bad because of the price, but not for Iphone 5S (assuming price remained same). we care about this because we may even come to know that economy of the people in INDIA increased within these 2 years (not taking all the factors into account).

Program Plan:

The project will be divided into the following four main components:

Each component is given weights. We do this because if two methods give different opinions, we go with the opinion given by the method which has more weight.

1. *Analyzing the emoticons*:

Emoticons play a key role in conveying information between users in the social networking sites. With the analysis of the emoticons choice of a user's tweet we can determine his feeling about a situation/event.

2. *Using geo-location to location information*:

Every twitter user have the capability of including their location while posting a tweet. We use the above mentioned information to classify the people according to their country.

3. *Understanding their tweets*:

We intend to understand the tweet by analyzing the words they have used to express their feeling about the phone. Then we aim to understand the meaning of the sentences they have used. This will help us to analyze their feelings in a better way. As an example, If we look at a tweet that has the following information "Iphone sucks bigtime". People using the terms such as "sucks" tend to have a bad feeling about the Iphone but the tweet "Iphone rocks galaxy phone sucks" also make the tweet to fall into bad feeling category. This analysis can be improved if we can combine words such as "Iphone sucks" to determine the users opinion. So that we can categorize them more precisely

4. *Using picture features to understand their feeling:*

People who feel strongly about a situation/event tend to include pictures to express their feelings. With the help of SIFT algorithm we can get the features of the picture they posted and then compare it with the database of images to get the feeling of the person

Example:

This kind of image can indicate that person who posted the tweet is not impressed with the new iphone and can be categorized as feeling bad about the Iphone 5 release.

Analysis Part(Evaluation) :

We try to include all the above mentioned methods to determine the people feelings using the tweets posted by them. After we have the information and analysis results. We can have the percentage of people who felt good, bad or neutral about a situation/event. We intend to use the same analysis but for a new data-set. This new data-set is the tweets of people who have responded when a similar situation/event has occurred in the past. In our example we will see if people felt good, bad or neutral when Iphone 4S is released(compared to data-set for Iphone 5S). We now have the percentage information of the two data-sets. With this information we can determine if the people opinion in total, changed or not. We can also determine the same for different countries and see how people are reacting to the situation/events now compared to past.

Plan of Action: (Rough Timeline)

1. First step is going to be collection of data-sets from twitter that has location information.
2. With the help of emoticons determining the users opinion.
3. Try to understand the terms used by the user and determining user's opinion.
4. If they have uploaded/used a picture along with their tweet. Using SIFT algorithm to determine the feeling of the user.
   (3-4 weeks to this point)
5. Categorizing the users according to their geological locations (countries).
6. According to the weights of the methods. We analyze the users opinions and classify them as good, bad or neutral.
7. Use the same method for new dataset, comparing the results
   (4 weeks to this point)

Referred Papers:

PAPER 1:

One of the paper that is basis for our proposal is **Sentimentor: Sentiment Analysis of Twitter Data** by James Spencer and Gulden Uchyigit. This paper uses a tool called Sentimentor, which is used for sentiment analysis of twitter data. And also how to automatically collect corpus for sentiment analysis and opinion mining purposes. The paper also uses Bayes classifier to classify the tweets considered into positive, negative or objective sets. This collect twitter API data and use a process called Tokenization, where it removes URL's, letters that are repeated more than twice continuously in a word and usernames that follow '@' symbol. It also removes the stopset data such as 'a', 'an', 'the' etc. which has no meaning of their own. They also considered POS tags. And finally their results are good when considering bigram without POS tags.

Here in our paper, we can use this sentimentor tool which uses the Bayes Classification to help us find positive, negative and neutral comments. We can also use this tokenization process if possible as to get the words from tweet that help us find the mood of the author regarding the situation/event. Considering all the process from the above paper that are applicable to ours and get efficient results on the opinion of the author.

This paper using data and text mining techniques and also large-scale information management which is related to our course work.

The strength's in this paper is are the bigram and POS tags used help us get more efficient results and weakness that we found are, they only worked on words in the tweet, which can be sometimes misleading when the user is sarcastic etc., their results also show there are atleast 45% false positives. Here in our paper we are working on using the words, emoticons and also pictures that are tagged so as good get results on finding whether the tweet is positive, negative or neutral.

PAPER 2:

The other paper that is basis for our proposal is **Emoticon Style: Interpreting Differences in Emoticons Across Cultures.** In this paper they tried to examine the use of emoticons on twitter and the variation of their usage across different cultures. Like in english speaking (western) countries, people mostly use horizontal type of emoticons and in eastern countries, people mostly use vertical type of emoticons. And also they determined the important factor affecting the style of emoticons is language rather than geography. They used **LIWC (**Linguistic Inquiry and Word

Count**)** program to that counts words in various psychological categories.

In our paper, we are trying to find the opinion of people on a situation/event basing on the emoticons they use in the tweet about that situation/event and comparing them basing on location of the post. Also we can use the **LIWC** program to get the terms used in those tweets as they also help us a lot in understanding the opinion of the author. Along with the words used in the tweet and emoticons, we also use the tagged pictures to determine the view of the author and also trying to understand the view expressed in post by grouping them based on geo-location, which gives us the idea whether people are happy,sad or neutral about the situation or event. The above paper also states there are variant forms for a normative form of emoticon but here we take the variant forms of a normative happy emoticon as happy and similarly for sad and neutral variant forms.

This paper deals with large-scale information management, data and text-mining techniques which is related to the course work.

The strength is above paper is they are also taking into account many kind of emoticons used across different cultures and languages to find the whether they are sarcastic or happy or sad in that tweet. The weakness is the when they are considering words with emoticons they are only taking 10,000 random which is very less when only taking emoticons and words into account. So here we are also using the tagged pictures which help us get the view author more efficiently.

PAPER 3:

The other paper that we have referred to is "**Sentiment Analysis of Twitter Data**" . In this paper the authors tried different methods to classify the user tweets including unigram model, a feature based model and a tree kernel based model. They have used emoticons, target and hashtags. Interestingly they have replaced words and acronyms to corresponding symbols using a dictionary. They designed a tree kernel model to evaluate the tweets. Finally they gave results for two classification tasks: 1) Positive versus Negative and 2) Positive versus Negative versus Neutral. This paper relates to the course work as it deals with large-scale information management. They even handled other languages using google translated tweets and a dictionary to understand the acronyms. They have tried five different methods and evaluated the results determining which methods results in better output. The weakness of the paper is that it does not take any other things into account to evaluate the results like parsing, punctuations , attached links etc..,

In the proposed project we try to enhance more on this paper taking more features into account to evaluate the results including picture matching, geo locating etc…,.

References:

1) **Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau**

Sentiment Analysis of Twitter Data

2) **James Spencer and Gulden Uchyigit**

Sentimentor: Sentiment Analysis of Twitter Data

3) **Jaram Park, Vladimir Barash, Clay Fink and Meeyoung Cha**

Emoticon Style: Interpreting Differences in Emoticons Across Cultures