# SENTIMENT ANALYSIS

CIE 553 – Natural Language Processing Course Project


Team Members:

Amr El-Gendy – 2013 0 48 26

Mohamed Maher – 2013 0 14 63

Table of Contents:

## Application Abstract:

Sentiment Analysis is the process of identifying, and categorizing opinions regarding a specific topic in order to determine the people attitude towards this topic whether positive, negative or neutral. This process is used widely to help in determining the marketing strategy, improving customer service.

In this course project, we are going to use the techniques learnt during the NLP course in addition what we have found in literature review to implement a sentiment analysis tool to overview people's opinions regarding any uploaded video on YouTube. This tool will be able to classify each comment in the video as a positive, negative or neutral opinion. Our application will target comment in English, Arabic, and Arabizi (Franko Arabic) languages. Moreover, the application will be implemented using python language and deployed on a cloud computing platform to be able to use it as a web application. So, the user will be able to just enter a YouTube video link then a summary of classified comments will be generated in a web page in text and graph format.

## Team Members:

1) Amr El-Gendy: 2013 0 48 26

   Roles:

   - Using YouTube API (4/5)
   - Data Preprocessing (Language Detection/Conversion – Tokenization and Normalization) (21/4)
   - Building Lexicon Approach. (28/4)

2) Mohamed Maher: 2013 0 14 63

   Roles:

   - Supervised approach by Building a Model from Labeled Datasets (21/4)
   - Application Integration and Collecting results (30/4)
   - Application Deployment (5/5)

## Application Technique:

In this application, we are going to use various techniques used in the NLP course in addition to ideas we have found in literature review.

Firstly, the gathering of data process will include connection with YouTube API in order to retrieve comments available in a certain video. Then, pre-processing steps will take place that will include various techniques like language detection, word normalization, tokenization, Franko-Arab Conversion and removal of stop words. After that, according to literature review found, we will implement two different supervised techniques in order to classify a comment:

1) Lexicon Approach: Using a lexicon of positive and negative words will help us to determine a certain weight of the opinion (i.e: each token will take a weight of +1, -1, 0 according to its value from the lexicon) then we will normalize total weight of the comment after adding values of all its token to determine whether it can be considered as positive comment (weight > 0.5) – Negative Comment ( Weight < -0.5) Neutral Comment (-0.1 < Weight < 0.1) or Uncertain (Weight has otherwise value).
2) Machine Learning Model: which will use a previous labeled dataset of opinions and using different machine learning algorithms (Bayesian Classifier, SVM or Neural Networks), we will be able to build a model that can be used in uncertain comments to classify them.

We have merged both techniques in our classification to ensure a high classification accuracy of the comments.

Finally, we are going to use python flask and Amazon web services to deploy the python code as a web application that can be used by different users by just entering a YouTube video hyper link to analyze its comments and classify them.

Note: The literature review showed us that dealing with Arabizi (Franko) language is still a crucial problem that hasn't achieved much progress in previous research. However, one of the solutions that we can implement is to use an API of any (Arabizi - Arabic) online translators like google translate to convert Arabizi comments to Arabic Language and deal with it in the same procedure as classification of Arabic comments.

**The main reasonable and insightful ideas in this course project can be summarized in the following points (Expected Output):**

- Using python to connect to a well-known API like YouTube in order to gather data needed for a NLP application
- Performing pre-processing steps on data obtained to prepare data for further processing
- Dealing with different languages and finding a solution to Arabizi Language
- Performing both supervised and unsupervised text classification in many NLP application process
- Learning about deploying of python scripts into a web application on cloud computing services like Amazon web services.

## Datasets used in course project:

1) **Machine Learning Approach:**

   A. “English Movie Reviews” in NLTK corpus

   B. “Aravec” Dataset

2) **Lexicon Approach:**

   A. “Hu and Liu” Lexicon for English words

   B. “NilULex” lexicon for Arabic words

## References:


1) AraVec Dataset: https://github.com/bakrianoo/aravec/tree/master/AraVec%201.0
2) 'Sentiment Analysis in Arabic: a Literature Review' made by Naaima Boudad in 2016 https://www.sciencedirect.com/science/article/pii/S2090447917300862
3) Youtube API Documentation https://developers.google.com/youtube/
4) GooglTrans API for using Google Translate https://pypi.python.org/pypi/googletrans
5) NileULex Lexicon (Arabic Lexicon Words) https://github.com/NileTMRG/NileULex
6) Hu and Lei (English Lexicon Words) https://github.com/woodrad/Twitter-SentimentMining/tree/master/Hu%20and%20Liu%20Sentiment%20Lexicon
7) Sentiment Analysis in Arabic tweets Rehab M. Duawiri in 2014, https://www.researchgate.net/publication/271550479_Sentiment_Analysis_in_Arabic_tweets
8) SENTIMENT ANALYSIS FOR ARABIC AND ENGLISH DATASETS By R.M. Elawady in 2015. https://www.researchgate.net/publication/281276441_SENTIMENT_ANALYSIS_FOR_ARABIC_AND_ENGLISH_DATASETS
9) Arabic / English Sentiment Analysis: An Empirical Study By Mohammed K. Alkabi et al. in 2013, https://www.researchgate.net/publication/236833290_Arabic_English_Sentiment_Analysis_An_Empirical_Study