YOUTUBE AUTOMATIC SENTIMENT ANALYSIS (YASA)

Natural Language Processing Project

Amr ElGendy

Mohamed Maher

Table of Contents

Abstract

Sentiment Analysis is the process of identifying and extracting relevant subjective information. It’s categorized as one of the difficult tasks of NLP and thus different approaches are being applied. In this project, we attempt to apply the concept of sentiment analysis to evaluate YouTube videos based on the analysis of the viewers’ comments (Arabic, English and Arabizi languages supported). To that end, we use YouTube API to extract the comments, supervised learning and lexicon based unsupervised learning for the evaluation. The application is then deployed as a web application service.

Introduction

Sentiment Analysis is considered as one of the difficult NLP tasks that needs special treatment and characterized by several methodologies. According to literature review, we have found that sentiment analysis can be done using un-supervised learning based on lexicon approach by having a large lexicon of used words in language with polarity of each word to help in determining the overall polarity of the comment. On the other hand, the supervised learning method based on using a machine learning algorithm on large data set of comments which helps in making decision of each new comment to be either positive, negative or neutral. However, in a research article entitled 'Sentiment Analysis in Arabic: a Literature Review' made by Naaima Boudad in 2016[2]. We have found new methods that combine both supervised and non-supervised approaches which are called semi-supervised approaches and we have used one of them in our analysis.

Methodology

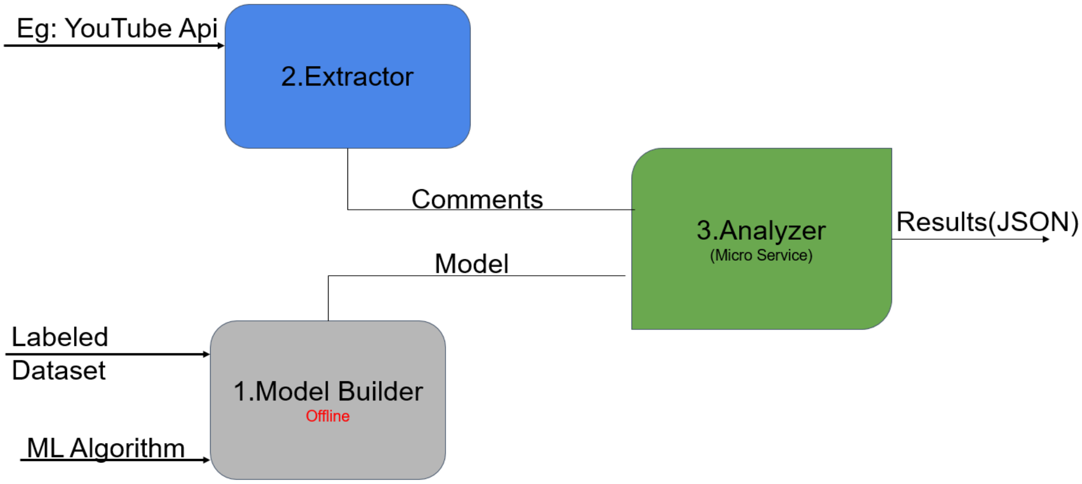


Figure 1 Summarized Procedure Chart

The whole process can be summarized as shown in Figure 1, these steps represent the milestones of the whole procedure. The detailed process is explained as follows:

I. Retrieving YouTube video Comments

The Algorithm starts by using YouTube API [1] in order to retrieve the relevant video information needed, namely the comments. This is achieved by an http request from the application to the YouTube API containing the ID of the video under investigation and the type of information requested (Comments in this case). The response of such a request is received in the form of a json file where individual comments are then extracted from and further analyzed.

II. Language Detection

As previously mentioned, the application supports 3 languages, namely Arabic, English and Arabizi. Detecting the target language in the comment investigated was chosen to be done by GoogleTrans API since it distinguishes Arabic and Arabizi languages, however, using the API requires the comment to be utf-8 encoded and without emotions. Thus before using the API, emotions are removed and the required format is applied. Arabizi language, once detected, is handled by converting it to Arabic using the API.

III. Preprocessing

For accurate results and optimum performance, first, stop words are removed from the comments, followed by tokenization of the comments. In addition to the preprocessing mentioned common to both languages, Arabic and English, Arabic is further preprocessed according to hand written rules, which includes converting all ‘أ’ or ‘إ’ to letter ‘ا’. Also, change all ‘ة’ to ‘ه’ and ‘ى’ to ‘ي’.

Following this step, one of two approaches is adopted depending on the situation.

IV. Lexicon Approach

This approach is used as a default one and is replaced with the second approach whenever it’s deemed insufficient. This approach is based on comparing the polarity of words in the comment under investigation to a labeled lexicon where each word is given a certain weight (E.g. love = 1, hate = -1). The total weight of each comment is calculated from tokens of the comment and is normalized so that its value ranges from +1 to -1 to detect the semantic of the comment. If weight of comment > |0.35| or less than |0.1| then we are almost sure of its class, otherwise, the second approach (building a model from labeled datasets) is used instead. “Hu and Liu” Lexicon is used for English words while “NilULex” lexicon is used for Arabic words. In the case of having words appearing in the lexicon of the respective language with multiple polarities (E.g. الله عليك – الله يخرب بيتك) the weight of the word is calculated by the following formula: Weight of word = (No. of times as Positive word in Lexicon – No. of times as Negative) / Total number of existence in lexicon. In addition, words preceded by negative words are assigned inverse weights (E.g. I **never** like such movies).

V. Pre-Trained Model Approach

In the case of the lexicon approach not being sufficient to analyze the comments (The weights being mid-range (|0.1| < polarity < |0.35|), this approach is used instead. This approach is purely a supervised learning one, where a dataset for each language (English Movie Reviews corpus in NLTK and Emotional Tone dataset for Arabic) is used to train a previously prepared model, the datasets are normalized and shuffled, and then the most frequent 10k words are chosen to represent features of the dataset. Then, it is divided into 80% training and 20% testing. Three Models were tested, deep neural networks (5 hidden layers and initial states of 1 and learning rate of 1e-5), Naïve Bayes and SVM. The best result was achieved by the DNN with accuracy of 99% for English language dataset and 82% for Arabic language dataset, therefore the other models were disregarded. The process of predicting the label of the comment starts by normalizing the preprocessed tokenized comments. Normalization in Arabic language was done by ISRIStemmer, one of the stemming tools included in the NLTK package. As for English, Porter 2 stemmer, a stemmer tool that proved to provide decent performance, was used. Features are then extracted from the normalized tokens and passed to the model for classification.

VI. Deployment

The python is deployed on Amazon cloud architecture as a web app and can be accessed from the following http://yasa-nlp.ml/ as in figure 2. The results are visualized using Google charts.

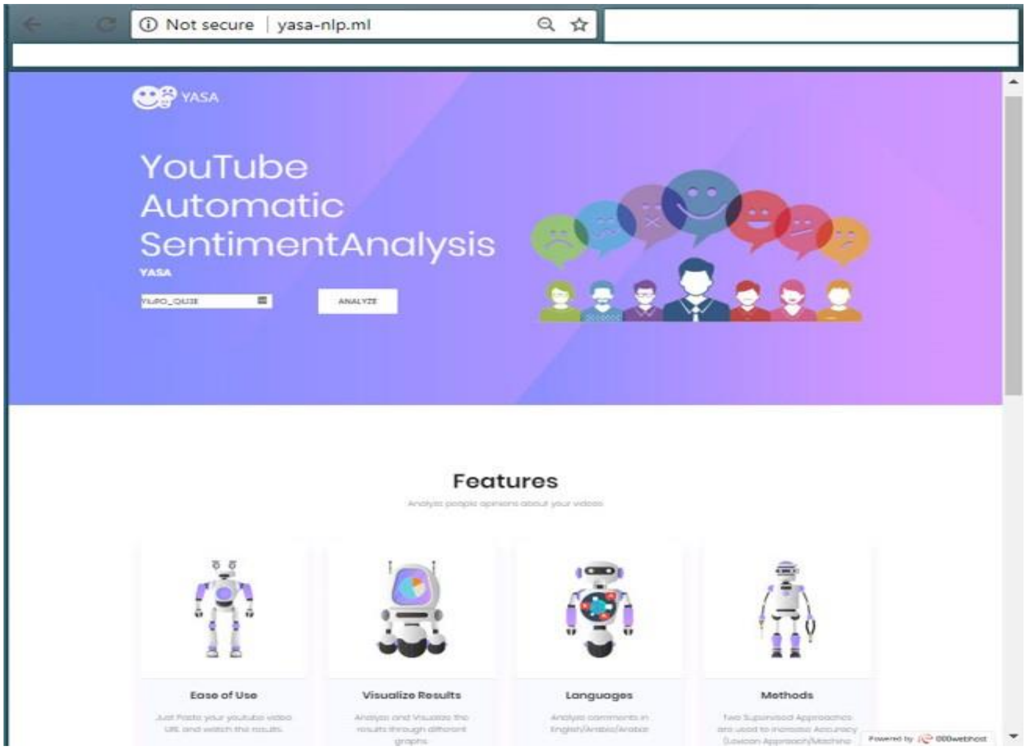

Not secure | yasa-nlp.ml
YASA
YouTube Automatic SentimentAnalysis
YASA
YLPO_QU3E
ANALYZE
Features
Analyze people opinions about your videos
Ease of Use
Just Paste your youtube video URL and watch the results.
Visualize Results
Analyze and Visualize the results through different graphs
Languages
Analyze comments in English/Arabic/Arabizi
Methods
Two Supervised Approaches are used to increase Accuracy (Lexicon Approach/Machine
Powered by 000webhost

Figure 2 A screenshot from the app in action

VII. Tools And Programming languages Used

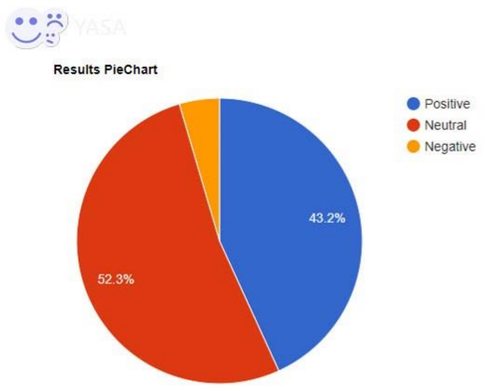
- Python and Python flask: classifier and model builder
- HTML, PHP, CSS, JS: creating webpages and communicating between the different APIs used,
- GoogleTrans API: for language detection and Arabizi to Arabic conversion.
- YouTube API: Acquiring YouTube video comments.
- Google Charts: Visualization of the results.

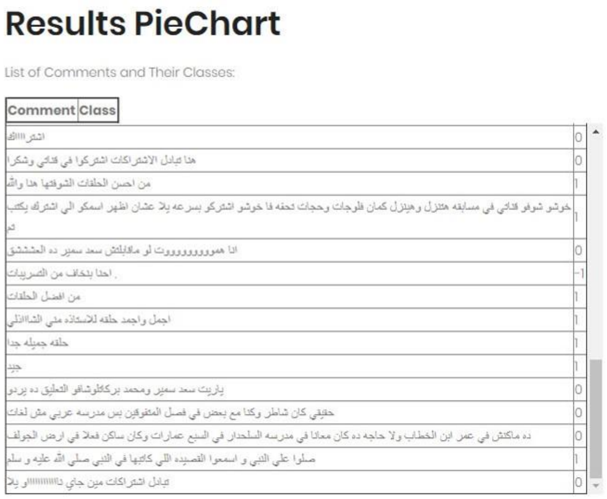
VIII. Results

The application was tested on several YouTube videos of different themes and trends. The following figures show the results of these tests and how they were visualized and presented.

Video 1 Test Results:

From the figures below, it can be observed that the first video had a slightly above 50% of the comments being neutral, with the rest being positive and almost negligible negative feedback. This accounts for a positive review for the video under investigation.





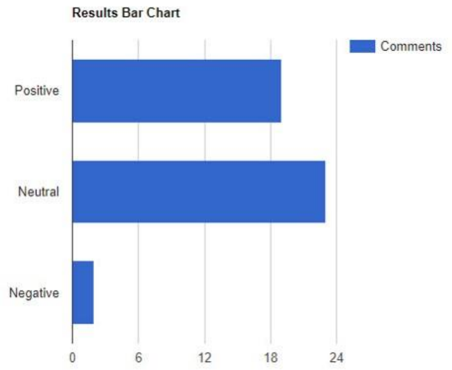
Results PieChart

List of Comments and Their Classes:

Comment	Class
اشترااااك	0
هذا تبادل الاشتراكات اشتركوا في قناتي وشكرا	0
من احسن الحلقات الشوفتها هنا والله	1
خوشو شوفو قناتي في مسابقه هتنزل وهينزل كمان فلوجات وحجات تحفه فا خوشو اشتركو بسرعه يلا عشان اظهر اسمكو الي اشترك يكتب تم	1
انا هموووووووووت لو ماقابلتش سعد سمير ده العشششق	0
احنا بنخاف من الضربات .	-1
من افضل الحلقات	1
اجمل واجمد حلقه للاستاذه مني الشاااذلي	1
حلقه جميله جدا	1
جيد	1
ياريت سعد سمير ومحمد بركاتلوشافو التعليق ده يردو	0
حقيقي كان شاطر وكنا مع بعض في فصل المتفوقين بس مدرسه عربي مش لغات	0
ده ماكنش في عمر ابن الخطاب ولا حاجه ده كان معانا في مدرسه السلحدار في السبع عمارات وكان ساكن فعلا في ارض الجولف	0
صلوا علي النبي و اسمعوا القصيده اللي كاتبها في النبي صلي الله عليه و سلم	1
تبادل اشتراكات مين جاي دااااااااا و يلا	0

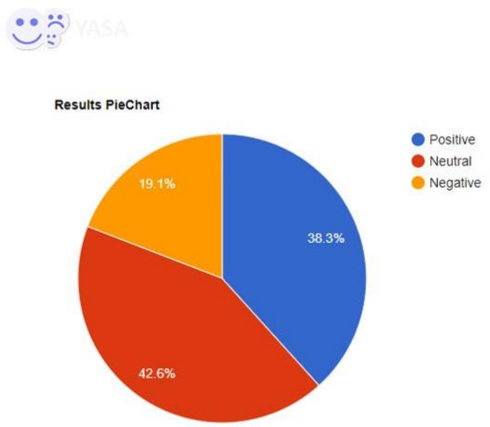
Results - BarChart

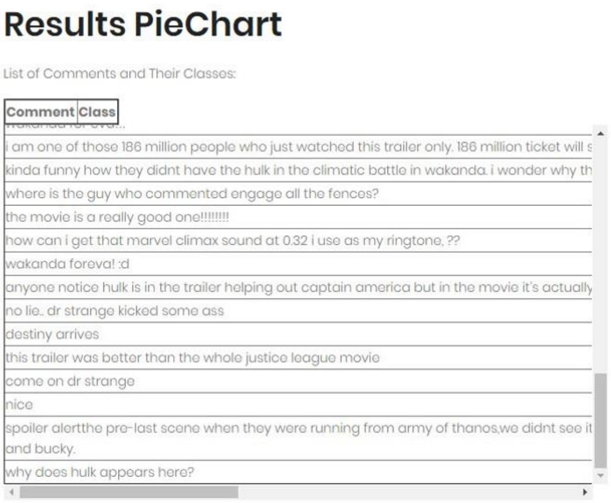
Representation of results of sentiment analysis in a bar chart



Video 2 Test Results:

The Figures below, representing the analysis of the second video, show a 19% negative feedback and a 38% positive feedback with the rest being neutral. This indicates how controversial the video reviewed seems to be.

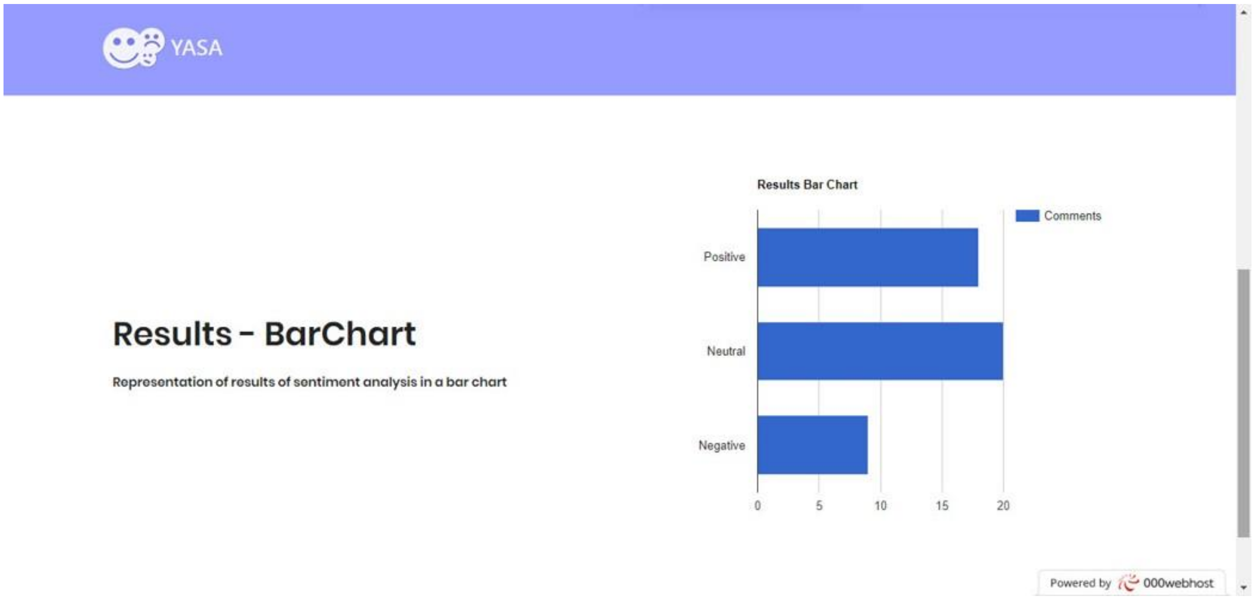




Results PieChart

List of Comments and Their Classes:

Comment	Class
wakanda for eva...	
i am one of those 186 million people who just watched this trailer only. 186 million ticket will s	
kinda funny how they didnt have the hulk in the climatic battle in wakanda. i wonder why th	
where is the guy who commented engage all the fences?	
the movie is a really good one!!!!!!!!	
how can i get that marvel climax sound at 0.32 i use as my ringtone, ??	
wakanda foreva! :d	
anyone notice hulk is in the trailer helping out captain america but in the movie it's actually	
no lie.. dr strange kicked some ass	
destiny arrives	
this trailer was better than the whole justice league movie	
come on dr strange	
nice	
spoiler alertthe pre-last scene when they were running from army of thanos,we didnt see it and bucky.	
why does hulk appears here?	



Results - BarChart

Representation of results of sentiment analysis in a bar chart

IX. Conclusion and Further Improvement

In conclusion, as explained, performing sentiment analysis is indeed one of the harder tasks of natural language processing and can be performed efficiently using the reliable supervised learning techniques such as DNN, SVM, and Naïve Bayes, achieving especially high accuracies in case of DNN, as well as the lexicon-based approach, which proved unreliable at times. Further improvements can be made internally by trying different and larger Lexicons, Using better Normalization techniques (Especially in Arabic), trying different ML algorithms with more parameters for Model Builder, Collecting Larger Datasets with wider domains for model builder with different type of features and trying different weight calculation approaches for Lexicon Approach. The domain of the application can also be expanded by widening its scope using different social media services other than YouTube and support more languages.

X. Appendix

File List can be found here

- Main Folder contains web page code for the application with styles folders to extract comments from YouTube API and display results using PHP.
- Python Folder contains:

1) **NileULex2.csv:** Arabic Lexicon of words.
2) **Hu and Liu Lexicon.csv:** English Lexicon of words.
3) **Negation_words.txt:** Contains negation words.
4) **Stop_words.txt:** Contains the stop words.
5) **Arabic_features / English Features:** features extracted using most frequent words in datasets.
6) **dnn_classifier_arabic.pkl / dnn_classifier_english.pkl:** models made by deep neural network to be ready to be used.
7) **Home.**py: contains python flask code of API used for sentiment Analysis.
8) **ModelBuilder.**ipynb: contains ipython code for building model offline.

References

1) Emotional-Tone-Dataset: https://github.com/AmrMehasseb/Emotional-Tone/blob/master/Emotional-Tone-Dataset.csv

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 https://github.com/NileTMRG/NileULex

6) Hu and Lei English Lexicon 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