CSE4022 – Natural Language Processing

PROJECT REPORT

Opinion Extraction from Online Review Videos

Submitted to

Prof. Arivoli A, SCOPE

Slot: E1+TE1

Submitted by

15BCE0076 - Vignesh Vaidyanathan

15BCE0082 - Voleti Ravi

15BCE0093 - Samudra Pratim Borkakoti

School of Computer Science and Engineering

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CONTENTS

1. ABSTRACT

Many users all over the world make use of Internet for sharing their experiences and giving opinions regarding a particular product or service on the World Wide Web and this phenomenon has increased the blogosphere. Blogs, which are also known as web logs are becoming an emerging source of information correspondingly increasing the interest in blog mining technique. Blog mining has many significant applications, like collecting opinions, reviewing search engine applications, etc. It is used for collecting and analyzing data. We have proposed a facet oriented scheme that is used to analyze the reviews given for a product and then assigning a sentiment label on each facet. There are score given to multiple reviews which are then aggregated and a comprehensive profile of a movie is produced on all the parameters. SentiWordNet based technique is used with two distinguished feature selection.

There are several challenges while mining the opinions from videos. For instance, to get the appropriate reviews about the product, opinion mining process should separate the reviewed data from the non-reviewed data. With the increasing needs of human beings, there is always an exponential increase in gadgets and other products. Although there are a variety of them, one has to make a decision on which product to buy according to its pros and cons. In this busy life, one cannot afford to go through all the specifications and compare the products.

In this project, the software goes through various review videos on a certain product in the Internet. The software converts the audio using speech-to-text and it uses the text to rate the product using sentiment analysis. This software can analyses any number of product review videos and calculate the overall rating for the product.

2. INTRODUCTION

Recent evidence suggests that some challenges and limitations exist in physical and online synchronous learning environments that still require attention to solve them. For example, on an academic event, information is usually addressed through audio channels so that students with learning or physical disabilities, foreign students, and other at risk populations are challenged to understand the content. Furthermore, one of the most common concerns reported in relation to online learning literature is the poor audio quality due to restricted internet bandwidth availability and traffic congestion. These problems can hinder students' understanding of a delivered speech, and this may hamper students from engaging in classroom participation and interaction.

According to related literature, abovementioned problems can be solved by adopting some assistive media-to-text recognition technologies, such as writing-to-text, image-to-text, diagram-to-text, text-to-speech, speech-to-text, and handwriting-to-text. For example, Speech-to-Text Recognition (STR) technology synchronously transcribes text streams from speech input and shows them on a whiteboard or students' computer screen. Furthermore, it is argued that STR-generated text can be employed as an additional text confirmation of what is being said.

2.1. Software requirements

- Python
- NLTK
- Jupyter
- Sphinx
- Numpy
- Audio files

3. LITERATURE REVIEW

The authors Jian Liu, et al. earlier have proposed and described an application on sentiment classification along with review extraction. This method works by extracting the review expressions on specific topic and by linking a sentiment tag and weight to the review expressions. After the process of tagging, a sentiment indicator of each tag is calculated by gathering the weights of all expressions. A classifier is used to predict the sentiment label of the text. The authors use online documents to test the performance of the proposed application.

A new method related to opinion mining is described in. The system is used to collect the positive and negative reviews.

The goal of the system is to extricate and condense the opinions and reviews, and then determining the reviews as positive or negative. The entire process is divided into four subtasks: content-value pair identification, expression identification, opinion determination, and sentiment analysis.

4. SYSTEM ARCHITECTURE

Initially we extract the audio from the video files of a reviews of a particular product and save them. We then further convert the audio file to text using a library package Sphinx which helps to transcribe the audio. We then apply the opinion mining on those text to obtain a score for each classification which is custom made and then o holistic score is given determining the overall review of the product.

4.1. Problem Definition

Opinions are reflected by the unprocessed and un-indexed the Web logs are filled with these texts. There are many people who make their decisions by taking other people opinions. For example, while buying a product, the customer will buy the product that is highly recommended by other people.

4.2. System Approach

Opinion mining (OM – also known as "sentiment classification") is a recent sub discipline at the crossroads of information retrieval and computational linguistics which is concerned not with the topic a text is about, but with the opinion it expresses. Opinion driven content management has several important applications, such as determining critics' opinions about a given product by classifying online product reviews, or tracking the shifting attitudes of the general public towards a political candidate by mining online forums or blogs.

1. **Determining text SO-polarity**, as in deciding whether a given text has a factual nature i.e. describes a given situation or event, without expressing a positive or a negative opinion on it or expresses an opinion on its subject matter. This amounts to performing binary text categorization under categories Subjective and Objective.
2. **Determining text PN-polarity**, as in deciding if a given Subjective text expresses a Positive or a Negative opinion on its subject matter.
3. ***Determining the strength of text PN-polarity***, as in deciding whether the Positive opinion expressed by a text on its subject matter is Weakly Positive, Mildly Positive, or Strongly Positive.

We achieve all this by using the **NLTK** (Natural Language Tool Kit), which is a library package with functions to classify the good opinions from bad opinions.

4.3. CMU Sphinx

The CMU Sphinx Team's tools is presented since they focus is speech recognition and it is the core part of the application. CMU Sphinx is a collection of several incarnations of Sphinx, a versatile continuous speech recognition tool-kit. We are more specifically using Sphinx-4 tool-kit. Sphinx-4 is an open source project. Sphinx-4 is modular and pluggable framework that incorporates design patterns from existing systems, with sufficient flexibility to support emerging areas of research interest. It supports any acoustic model structure and supports most of the language models.

The Sphinx-4 has three principal modules, the FrontEnd, the Decoder and the Linguist getting material of the Knowledge Base.

- The FrontEnd gets a single or several input signals and computes them so that a sequence of Features (vectors) is created.
- The Linguist generates a SearchGraph by translating any kind of standard language, with the aid of pronunciation information contained in a Lexicon, and with the structural information stored in sets of AcousticModel.
- The SearchManager component located in the Decoder uses the Features from the FrontEnd and the SearchGraph from the Linguist to perform the actual decoding, generating Results. During or prior to the recognition process, the application can issue Controls to each of the modules.

5. METHODOLOGY

Here we ask the user to select the product they want to be reviewed. The product is then looked for in YouTube, and the 5 most relevant videos are selected. The audio files for these videos are extracted, which are converted into text files using Sphinx Voice Transcriber and stored in folders representing the product.

Using a custom classifier, we have classified each review of the product into different specifications – Build quality, Processor, Display, Camera, Battery, Audio and so on - and put them individual files.

We then use the NLTK to determine the polarity of individual sentences as a compound. As per NLTK trainer the final sentiment is determined by the classification possibilities below:

Polarity of statement:

• Positive

• Negative

• Neutral

There are three numerical scores associated with WordNet synsets. These scores are Obj(s), Pos(s) and Neg(s), that describes how objectives, positive and negative terms are contained in the synsets.

To evaluate the performance of the proposed application, we used several reviews on each product from the file system. For our analysis we simply chose latest and trending smartphones, since we want to analyze as many verifiable reviews as possible.

As per given information we also calculate rating out of 10 using following formula. The result will be calculated as per the formula given below with positive and negative scores generated by the classifier. This score is made to be graded out of 10.

Here, **compound = ∑ (positive + negative)**

The final compound score generated is normalized to a rate out of 10 by

Normalized score = 5*compound + 5

User can see detailed reviews of the desired product of YouTube videos in textual and graphical format. A line graph is generated to show final result of total reviews visually.

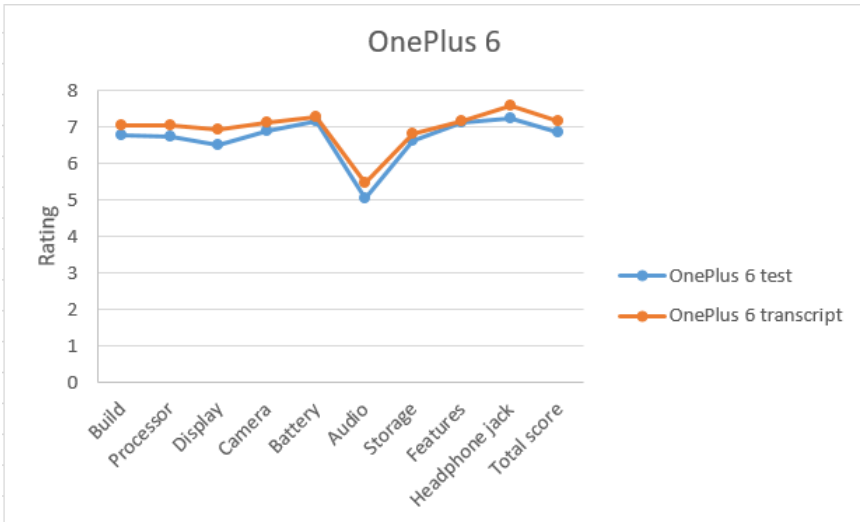
6. RESULTS

```
[REVIEW] PRODUCT
Build_quality =  6.77628378378
Processor =  6.751625
Display =  6.51879411765
Camera =  6.91793150685
Battery =  7.1578
Audio =  5.05783333333
Storage =  6.6148125
Features =  7.14676595745
Headphone_jack =  7.24751724138
PRODUCT SCORE =  6.84243032993
--------------------------------------------------------------------------------
BEEP-BEEP ANALYSIS COMPLETE BEEP-BEEP
--------------------------------------------------------------------------------
xelese@xelese:~/NLP$
```

Fig1. Scoring of OnePlus 6 test

```
[REVIEW] PRODUCT
Build_quality =  7.14755555556
Processor =  7.05914909639
Display =  6.99399765258
Camera =  7.05858169014
Battery =  7.23132727273
Audio =  5.65827380952
Storage =  6.73955737705
Features =  7.12703266332
Headphone_jack =  7.53580869565
PRODUCT SCORE =  7.15901853806
--------------------------------------------------------------------------------
BEEP-BEEP ANALYSIS COMPLETE BEEP-BEEP
--------------------------------------------------------------------------------
xelese@xelese:~/NLP$
```

Fig2. Scoring of OnePlus 6 transcript

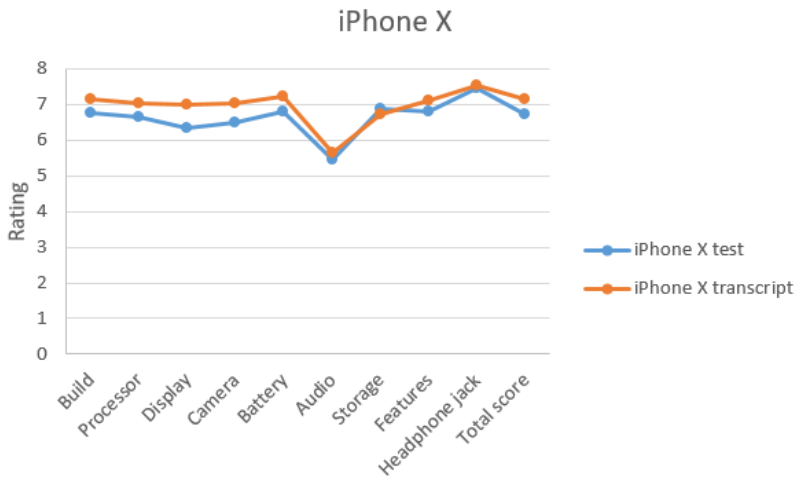


```
[REVIEW] PRODUCT
Build_quality =  6.77281
Processor =  6.63538562092
Display =  6.3428172043
Camera =  6.51376923077
Battery =  6.80613157895
Audio =  5.47792857143
Storage =  6.86066666667
Features =  6.80413235294
Headphone_jack =  7.47021111111
PRODUCT SCORE =  6.74699863342
--------------------------------------------------------------------------
BEEP-BEEP ANALYSIS COMPLETE BEEP-BEEP
--------------------------------------------------------------------------
xelese@xelese:~/NLP$
```

Fig3. Scoring of iPhone X test

```
[REVIEW] PRODUCT
Build_quality =  7.05531140351
Processor =  7.04245940959
Display =  6.95502409639
Camera =  7.14484302326
Battery =  7.28590909091
Audio =  5.49142307692
Storage =  6.83191525424
Features =  7.18308536585
Headphone_jack =  7.59248421053
PRODUCT SCORE =  7.15802442865
--------------------------------------------------------------------------
BEEP-BEEP ANALYSIS COMPLETE BEEP-BEEP
--------------------------------------------------------------------------
xelese@xelese:~/NLP$
```

Fig4. Scoring of iPhone X transcript



7. CONCLUSION AND FUTURE SCOPE

Opinion mining is an interesting area of research. As World Wide Web increases it generates huge amount of information, extracting such information has become an important task. In this project we proposed a web mining application which is used for calculating scores from web blogs or any text obtained and display their review scores. We have implement and obtained what was proposed and was expected as a result respectively. We used web crawling, Sentimental analysis via NLTK and custom classification and visualization approach to get the expected result.

For the future study, we can improve this application by adding extra features like Spell Check and movie rating system to verify manipulated reviews for improving the performance and accuracy and also plan to explore more problems related to NLP issues in the future work.

8. REFERENCES

[1] Qiang Ye, et al., Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, Expert Systems with Applications (2008) doi:10.1016/j.eswa.2008.07.035

[2] Classification and Summarization on rating of Mobiles features, Pallavi Bharambe, Prof. Sanjivani Deokar, and Department of Computer Engineering. Lokmanya Tilak college of Engineering, Navi-Mumbai, India

[3] Review of Speech-to-Text Recognition Technology for Enhancing Learning Rustam Shadiev, Wu-Yuin Hwang, Nian-Shing Chen and Yueh-Min Huang, Journal of Educational Technology & Society, Vol. 17, No. 4, Review Articles in Educational Technology (October 2014), pp. 65-84