**Literature Review: Youtube Data Analysis**

Bowen Feng Howard Hu Jing Lin Yuchen Zhao

**Teammate who read and summarized**: Howard Hu

**Paper:** Lim, Ji Young; Kim, Seulki; Kim, Juhang ; Lee, Seunghwan. Identifying trends in nursing start-ups using text mining of YouTube content. 2/13/2020, *PLoS ONE*,Vol. 15 Issue 2, p1-14. 14p. ([link](http://web.b.ebscohost.com.libproxy.rose-hulman.edu/ehost/pdfviewer/pdfviewer?vid=6&sid=9a880dc5-c939-4f0e-bbe2-3986dd5f6b3d%40sessionmgr101))

**Summary:** Authors illustrate that using YouTube content can analyze top trending videos in different areas. There are three main steps in this analysis. They are text mining, Delphi survey, and comparison. For text mining, they could analyze frequency and degree centrality of different keywords in trending videos. For the Delphi survey, they use a synthesis framework for using start-up to classified about one hundred frequent keywords. For the comparison, they use two times two matrix to fit the results of the previous two steps. Authors conclude that text mining helps them to identify some keywords such as “area”, “business”, and “competence.” The Delphi survey shows that the highest level of frequent keywords are “motivation”, “advice”, “obstacle” and so on. In the comparison step, “dream”, “idea”, and “opportunity ” are top keywords in the two times two matrix. Authors conclude that it is important for people to develop business skills such as finance and budgeting in the early stage of nursing start-ups, and active policy helps people to prepare for the Fourth Industrial Revolution.

**Applicability:** Since our project is related to a lot of natural language processes, we could use the keyword frequency test to figure out which keyword is the most attractive word in the video titles. The Delphi survey is a good way to help us to find the most frequent keywords in the early stage of data analysis. This method gives a rough prediction about keywords. Then we could dig deeper into it like generating a machine learning model.

**Issues:** Authors did not generate a machine learning model based on their data. They have a comprehensive analysis of keywords, but keywords are not the only influential element in trending videos. Time, location, and video length can be other important influence elements of a top trending video. Therefore it is necessary to include every possible element when we analyze the trending video. Their data set is not large enough. They focus on one specific area like business area with limited data. Then they may have some bias and variance in predicting the most frequency keyword.

**Teammate who read and summarized**: Jing Lin

**Paper:** Shaila S.G, Prasanna MSM, and Kishore Mohit. Classification of YouTube Data based on Sentiment Analysis. *IJERCSE*, Vol 5, Issue 6, June 2018. ([link](https://www.technoarete.org/common_abstract/pdf/IJERCSE/v5/i6/Ext_92830.pdf))

**Summary:** Overall, the authors create a model that helps them to study the sentiment from the consumer’s perspective based on classifying the youtube topics. One of the algorithms they use to classify the Youtube data based on user sentiment analysis is Linear regression, which we already covered in class. Before they have any conclusion, they use the corr() function and the heatmap to clearly show that the Views, Likes, Dislikes, Comments are correlated with about 85% chance, And the dot plots also show the linear relationship between Views and Likes. One thing surprised me is you can plot multiple pair plots with the code this paper provides. So we can analyze the relationship between features as a whole.

**Applicability:** Like Dr. Shibberu mentioned in class, we will spend most of our project time on preprocessing the data. One of the reason would be some data in our dataset might be redundant and has noisy values, which might impact our predicted result severely. This journal demonstrates a couple of ways to clean the data, which will give us some idea of what data we should get rid of. One way to check whether the dataset has the missing data is to build a heatmap that can visually show us whether the dataset has missing data because we can set the missing data with a different color.

**Issues:** This paper only shows how people can analyze the Youtube data, it does not actually provide a deep analysis of Youtube data. It only talks about the relationships among the features, it did not really talk about the reason behind the relationships. And I am sure there are some relationship that is more complex than the linear relationship, so using corr() and linear regression is not enough. It mentioned the Logistic Regression, unfortunately, it did not demonstrate the example of Logistic Regression.

**Teammate who read and summarized**: Yuchen Zhao

**Paper:** SZABO, GABOR, HUBERMAN, BERNARDO A, Predicting the Popularity of Online Content, *Communications of the ACM,* Aug2010, Vol. 53 Issue 8, p80-88 ([link](http://web.a.ebscohost.com.libproxy.rose-hulman.edu/ehost/pdfviewer/pdfviewer?vid=1&sid=a79eb719-2675-4db0-9e83-e240f7b7db31%40sdc-v-sessmgr03))

**Summary:** This article shows the methods of collecting a dataset and predicting the popularity of the news and videos. It mainly illustrates the data collecting method and predicting method. As the advantages and disadvantages are being discussed, the authors explain the limitation of predicting future popularity with the existing dataset. In conclusion, the article mentioned that the prediction is rather limited due to observations from existing data. It does not have the ability to see why something becomes popular.

**Applicability:** Among all the aspects being discussed in the article, the most valuable point is how we should avoid misleading data. Like the author discussed in the article, the provided dataset is updated daily, which may not be enough for a precise prediction. However, for our project, we are aiming to accomplish a rather vague prediction of future trends. As the article provides what information in the dataset we should process, I would consider it as a good reference to start up the project.

As the project goes on, we need to develop more detailed algorithms to train the existing dataset. The articles use the comparison of Youtube dataset and Digg dataset to show how the training progress is useful in both dataset.

**Issues:** It has very limited resources about how to apply the data processing progress in detail. As we need to use the machine learning skills to finish the task, the article only provides a general idea rather than an approach. The relationships between the data has been briefly explained, but there is no further explanation. We can see it mention how the linear correlation exists, but there are very limited explanations about it.

I do not think the algorithms mentioned in the article are applicable in our case, but it is still a good reference.

**Teammate who read and summarized**: Bowen Feng

**Paper:** Predicting Success of Bollywood Movies Using Machine Learning Techniques

<https://dl.acm.org/doi/pdf/10.1145/3140107.3140126>

**Summary:** The authors collected the data from Youtube and other video channels to predict if the films can achieve success or not. The data including their trailers’ length and the rating of their actors. They used classification and feature engineering approach to do the analysis.

**Applicability:** For the nature of machine learning, we will spend most of our project time on preprocessing the data. When we have some data that is not numerical and need to be rated as factors to affect the prediction result. We need to pay attention on how to transfer some facts into data based on some authority sources.

**Issues:** This paper didn’t show detailed information of how they do the feature engineering or the classification. The real process can be hard than we see and may take longer.