University of Southampton

Faculty of Engineering and Physical Sciences

Electronics and Computer Science

Popularity and Traffic Value Prediction Based on YouTube Videos Comments

by

Yubo Zhang

September 2019

Supervisor: Dr Markus Brede

Second Examiner: Professor Timothy J Norman

MSc Artificial Intelligence

Abstract

With the development of Internet, people can learn different things happening in the world in quite a high speed. In the field of online video, people can watch videos of interest, post user comments or leave messages on major video sites. These comments can express the user's views and attitudes to a certain extent, so they are very valuable to video sites, and video producers. This project aims to explore an approach that can be used to analyze the feedback based on Natural Language Processing (NLP) technology according to YouTube comment data. It can recognize the attitude of audiences and predict the traffic value of a video.

Machine learning has been a popular technology of automatic feature extraction and application in Artificial Intelligence field. In this paper, the prediction instrument is researched and developed based on machine learning. According to current research basis, sentiment analysis module, regression analysis module and keyword extraction module are implemented in the instrument mainly based on word embedding, deep learning and tf-idf algorithm. By using YouTube dataset, the prediction system has been designed and evaluated combining sentiment analysis module and regression analysis module. The keyword extraction module is a supplementary part so that the comment analysis is intuitionistic and abundant. Finally, some discussion is given in order to research the topic further.

Abbreviations

NLP Natural Language Processing

NN Neural Network

Tf-idf Term Frequency–inverse Document Frequency

SVM Support Vector Method

RNN Recurrent Neural Network

CNN Convolutional Neural Network

LSTM Long Short-Term Memory Neural Network

API Application Programming Interface

Statement of Originality

- I have read and understood the ECS Academic Integrity information and the University's Academic Integrity Guidance for Students.
- I am aware that failure to act in accordance with the Regulations Governing Academic Integrity may lead to the imposition of penalties which, for the most serious cases, may include termination of programme.
- I consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to verify whether my work contains plagiarised material, and for quality assurance purposes.

You must change the statements in the boxes if you do not agree with them.

We expect you to acknowledge all sources of information (e.g. ideas, algorithms, data) using citations. You must also put quotation marks around any sections of text that you have copied without paraphrasing. If any figures or tables have been taken or modified from another source, you must explain this in the caption and cite the original source.

I have acknowledged all sources, and identified any content taken from elsewhere.

If you have used any code (e.g. open-source code), reference designs, or similar resources that have been produced by anyone else, you must list them in the box below. In the report, you must explain what was used and how it relates to the work you have done.

I have not used any resources produced by anyone else.

You can consult with module teaching staff/demonstrators, but you should not show anyone else your work (this includes uploading your work to publicly-accessible repositories e.g. Github, unless expressly permitted by the module leader), or help them to do theirs. For individual assignments, we expect you to work on your own. For group assignments, we expect that you work only with your allocated group. You must get permission in writing from the module teaching staff before you seek outside assistance, e.g. a proofreading service, and declare it here.

I did all the work myself, or with my allocated group, and have not helped anyone else.

We expect that you have not fabricated, modified or distorted any data, evidence, references, experimental results, or other material used or presented in the report. You must clearly describe your experiments and how the results were obtained, and include all data, source code and/or designs (either in the report, or submitted as a separate file) so that your results could be reproduced.

The material in the report is genuine, and I have included all my data/code/designs.

We expect that you have not previously submitted any part of this work for another assessment. You must get permission in writing from the module teaching staff before re-using any of your previously submitted work for this assessment.

I have not submitted any part of this work for another assessment.

If your work involved research/studies (including surveys) on human participants, their cells or data, or on animals, you must have been granted ethical approval before the work was carried out, and any experiments must have followed these requirements. You must give details of this in the report, and list the ethical approval reference number(s) in the box below.

My work did not involve human participants, their cells or data, or animals.

ECS Statement of Originality Template, updated August 2018, Alex Weddell aiofficer@ecs.soton.ac.uk

Content

List of Figures

Chapter 1 Introduction

1.1 Background

Due to the development of media and Internet technology, the online video market has been greatly prosperous in recent years. The network audio-visual content has become the spiritual consumption of online users, which causes economic benefits and public influence. The types of programs that users prefer to watching are obviously different due to gender, intergeneration, and regional culture. Since the amount of audience increasing, it is probably difficult for video publishers to analyse audiences' feedback and judge the effect of their works. Therefore, it is demanded for publishers to accurately analyse feedback of a video and predict its potential value for future activities such as sequels or series plans.

Generally, online users usually watch the videos which are related to their interested topics or which are uploaded by their favorite publishers, and then some of the viewers will post their comments. These comments have intrinsically related features, expressing audience's opinions and attitudes. An advanced approach that will be explored in this paper is achieving sentiment and textmining analysis using Natural Language Processing (NLP) technology. In order to refine this problem and provide a generally acceptable evaluation methods, YouTube is selected as the database provider to do the research. YouTube is a popular international online video platform. A large number of video producers share their works on it. As of 2019, There are over 1.3 billion users on YouTube and in an average month, 8 out of 10 18-49 year-olds watch YouTube. However, the quality of video is complex. Only attractive contents may be widely spread and bring profit to publishers. The starting point of the project is to use NLP technology to judge what is people's attitude towards the film, and then try to summarize a group of core reviews according to the comments. It is helpful for video publishers to recognize the degree how people like the video and conclude the

overall evaluation of the video.

1.2 Current Research Status

The popularity and traffic value prediction of online video has been a topic of Artificial Intelligence for several years. There are many methods explored to predict the view number of online videos, including SARIMA model, Deep Belief Networks, k-neighbor algorithm, etc. Most of the previous studies focus on time series issue. For instance, Yan Liu designed a prediction software based on SARIMA model. She designed the prediction model combining data extraction program and relative time series analysis theory of SARIMA model, which has ability to predict future statement of an online video with high correlation.Moreover, researches based on video cover or abstract text are also topical subjects. Taken together, the prediction task based on natural language processing algorithm is still a challenge area in this field.

Natural Language Processing technology has been a popular research area since early 2000. It was researched based on early theories of phonology, morphology, syntax, semantics and lexicography. With the development of computer science, the computing ability was greatly improved which was helpful to achieve complex NLP algorithms. Meanwhile, the technology was also developing from bag-of-words model, n-grams model to Word2Vec model. Nowadays a number of new models have been invented and applied to solve challenging topics, including sentiment analysis, document classification, machine translation, text summarization and question answering.

As for prediction or regression issue, the algorithms are diverse and multifunctional in machine learning field. Each machine learning algorithm has its own shortcomings, which also allows us to refer to the choice. For example, the Recurrent Neural Network (RNN) is a very popular model and has shown great performance in many of

machine learning tasks. Although an algorithm is not universal, each algorithm has some features that allow people to quickly select and adjust parameters. The key is to grasp different models and test their performance in different tasks.

1.3 Project Aims and Objectives

1.3.1 Project Aims

The project aims to explore the achievement of a text analysis instrument. The instrument should contain Artificial Intelligence algorithms so that is able to complete the prediction and analysis tasks. The prediction part is the main aim of this instrument, which describes the business value of the online video. The text analysis part focuses on key word extraction, which is a good supplemental tool that implies the focus points of the video. The goal of the prediction part is to generate the level of view number as output, based on the comment feedback. Besides, the goal of the text analysis part is to extract five effective words or phrases with highest frequently discussed in the comments. It is a good supplementary specification of video analysis about its hot spot.

Ideally, the accuracy needs to reach an acceptable level. As a different standpoint of solving the issue in online video traffic value prediction, the instrument can be a good supplementary method. The relative percentage of prediction error of current prediction systems can reach around 20% to 30%. Due to the subjectivity of comments text, the instrument in this project is hoped to reach lower than 40% of prediction error.

1.3.2 Project Objectives

First of all, in order to achieve the algorithm and train the model, the project needs to obtain a dataset. The dataset need to be huge enough so that it can be used to build a

reliable model. YouTube API is an official tool which can be used to comprehensively scrap the information of a video and then managed as original dataset. Video id, view number, comment text and other information need to be included and built as a .xls file.

Then, by employing python, the instrument can be experimented and completed. There are several necessary steps in order to process the original dataset. After pre-processing dataset, the comment text is prepared for machine learning using word embedding and sentiment analysis technology. According to Pozzi et al, their paper demonstrate the tendency of viewers' opinion influences the video's popularity, which gives theoretical support. I can use the processed dataset to build a regression model that have ability to predict the view number of videos. On one hand, word embedding in machine learning technology field is quite mature that there are several method to classify the sentiment of short text. On the other hand, because of the fact that the dataset is large, deep learning can be explored to solve the prediction issue. Different kind of neural network algorithms are tested, including CNN, RNN, etc. By comparing the result of different algorithms, the model with best performance will be selected as solution.

Furthermore, the key word extraction part is also designed by employing python. For each video, the comment text is collected to find five effective word which appear with highest frequency. Useless text data are filtered out and the main substance of comments is indicated.

1.4 Project Management

The project was completed in around 14 weeks. The following Gantt chart is the plan schedule of the project. The arrangement is scientific. At the beginning, I focused on study and research relative knowledge. Next I learned to use YouTube API and scraped the data. Then I processed dataset and designed the whole program part by part. Meanwhile I finished the experiment and debug the system. At last I completed the final report in the rest of time.

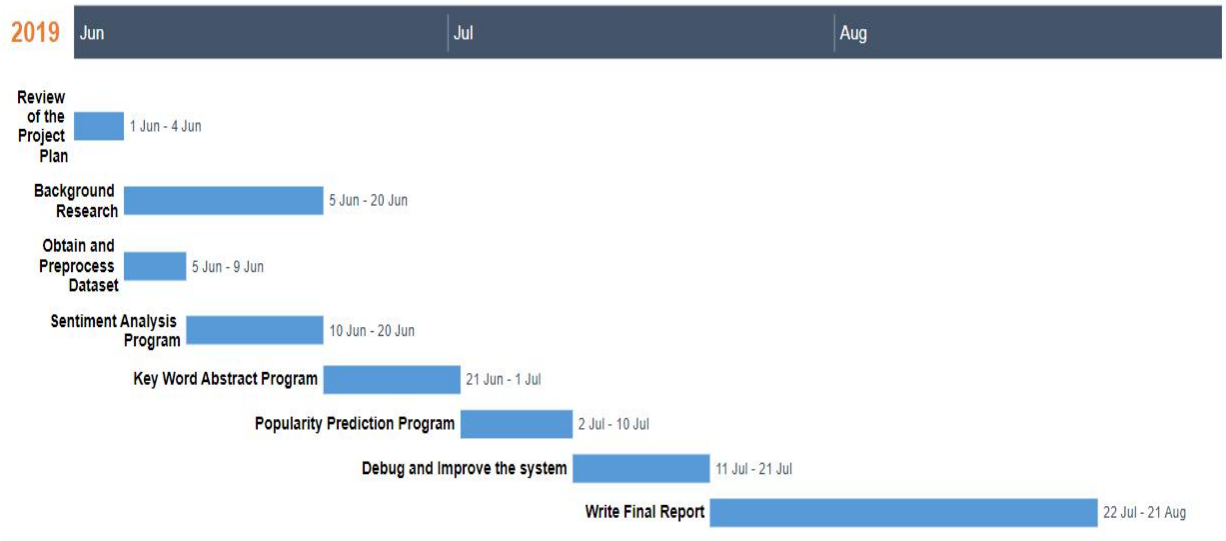

Figure 1 Project Gantt Chart

1.5 Report Structure

There are 6 chapters and one reference page in the report. A design archive (.zip file) including code and dataset is also attached.

Chapter 1 Introduction

The introduction chapter gives reader the background information and explain the project. The aim and objective of the project is also given. The schedule of the project is shown by a Gantt chart.

Chapter 2 Related Text Analysis Technology

In this chapter, an overview of related Artificial Intelligence sources are provided in literature review, including regression model, CNN model, RNN model with LSTM model, and key information extraction method. Moreover, an introduction of python and YouTube API is also given in “tools and applications employed in the project” section

Chapter 3 Problem Definition and Instrument Design

In this chapter, the problem is defined and explained in detail. I will clarify how to solve the problem in the project and how the system work.

Chapter 4 Experiments

According to the treatment of the problem above, I demonstrated how I implemented the theories when engaging in project. The process of designing and fitting the models, evaluating the results and generating key words will be shown in this chapter.

Chapter 5 Legal and Commercial aspects

In fact, a lot of public data was used in the experiment. The instrument is also designed to be practiced in a real data environment. Therefore, I will discuss the legal and commercial problem of this project.

Chapter 6 Conclusion

In this chapter, a discussion about the relation between real world and the project is included. I will also share my opinion of this project and what further research need to do.

Chapter 2 Related Text Analysis Technology

2.1 Literature Review

During the research, I combed the fundamental sequence of ideas that NLP and regression theories develop. All the Literature used for this dissertation includes academic textbooks, online documentation and IEEE research papers. There are approximately 20 literatures cited and recorded in reference part.

2.1.1 Sentiment Analysis

Sentiment analysis process can be defined as a typical classification process, i.e., input text and output positive or negative emotional categories. Some literatures put forward naive bayes or svm classifiers which can be often used for classification. There are even some opinion corpora that can give an objective sentiment vector for each word directly. The general classification algorithm is more reliable than some recently proposed algorithms that are dedicated to text analysis, but it is necessary to firstly convert the text into real vector as the input of the classifier. Therefore, the biggest unstable point of these algorithms in application comes from the modeling of eigenvector, that is, different modeling schemes will achieve very different results even if the same method is used. It can also be seen from the previous papers of sentiment analysis that a considerable number of sub-papers have made breakthrough in classification effect due to the adoption of new feature parameters.

The method of text classification using general classifiers has been proposed for a long time, and the method of applying it to sentiment analysis is firstly proposed in Pang B's paper. In this paper, the tf-idf method is used to transform the text into feature vector, which is the basis of many improved methods. The tf-idf method is based on the hypothesis of word bag (bag-of-word, BoW). The bag-of-word hypothesis ignores the relationship between words in the document and regards the document as a disordered set of words. Therefore, the tf-idf method determines the contribution of words to the document according to the frequency of words in the

document and the universality of words in the corpus, and takes this contribution value as a component of the document feature vector. Following I will simply introduce this algorithm. Firstly, let the total number of words be N_t. Therefore, each documents can be modeled as a N_t - dimensional feature vector . Then set the frequency of the statistic word in the document as $n_i(d)$, i.e.

$$\boldsymbol{f}(d) = \left(f_1(d), \dots, f_{N_T}(d)\right) \qquad i \in [1, N_T]$$

So the frequency at which the word appears in the document tf_i is:

$$tf_i(d) = \frac{n_i(d)}{\sum_{j=1}^{N_T} n_j(d)}$$

We can define hit function δ_i^{corpus} and calculate The number of times words appear in the corpus n_i^{corpus} :

$$\delta_i^{corpus}(d) = \begin{cases} 0, & n_i(d) = 0 \\ 1, & n_i(d) > 0 \end{cases}$$

$$n_i^{corpus} = \sum_{j=1}^{N_D} \delta_i^{corpus}(d_j)$$

Where N_D is the amount of document in the corpus. In order to facilitate the expression, the universal of the word df_i is the logarithm of the frequency of the word in the corpus, and the smoothing processing is also added to ensure the boundary of the data.

$$df_i = \log \frac{1 + n_i^{corpus}}{1 + N_D}$$

The contribution of words to the text is positively correlated with the frequency of words appearing in the text, but negatively correlated with the universality of word distribution. Therefore, defining the universal contribution value idf_i and final eigenvector component f_i as:

$$idf_i = -df_i$$

$$f_i(d) = tf_i(d) \times idf_i$$

Unlike tf-idf model, according to the paper written by Mudinas A, Zhang D and Levene M, an idea of Semantic algorithms are proposed, which are used to construct

the idea of eigenvectors and machine learning algorithms.This method is different from the tf-idf method. The literature firslut uses a list of sentiment words to search all sentiment words that appear in the corpus. Each word in the list of sentiment words corresponds to an emotional value defined in the interval [-3, +31 (manually defined, positive and negative representing emotional direction). The literature then considers that the emotional value ultimately expressed by the emotional word depends on the corresponding emotional value of the list word and the correction of the emotional value. The amendments are divided into two types: negative correction and correction correction. The negative corrections used in the paper include negative words ("not" and "n't") and negative imperatives ("stop doing" and "quit doing") that can be matched to emotional words. These matches set the emotional value to The opposite is true. The paper also adds a generic negative word feature for negatives that cannot match emotional words, and the emotional value is set to -1.5. As for double negation, the paper uses a negative phrase ("not only" and "notjust") to ensure that it is excluded from the negative. The correction function used in the paper also uses the word matching method, and the matching words include 75 words of comparative of adjective, comparative of adverb and adverbs of degree.

2.1.2 Regression Analysis

The Regression Analysis prediction method was originally proposed in a study of human genetic problems by Francis Galton, a famous British statistician. By observing the height data of 1078 couples and their children, he found that there is a linear relationship between their height and the heights of their parents, so that the height of the parents can be used to predict the height of their children.

Regression analysis is essentially a function estimation problem that to find the causal relationship between Dependent Variable (DV) and Independent Variable (IV) Then on the basis of correlation, the regression equation between variables can be

established. Regression analysis and forecasting is a common forecasting method that is simple, efficient, explanatory, and practical.

In the late 1980s, Barry Litman, the pioneer of film box office revenue forecasting, used regression analysis prediction model for the first time to analyze 697 films shown in the United States in 1980s. Because of the limitation of data acquisition, he regarded film rental income as the forecast value of ticket room volume, and concluded that the success of a film box office was determined by its special factors, i.e. the story, the release schedule and marketing work. Each of the key factors is divided into many sub-items which need to be further analyzed quantitatively by regression analysis to predict the success of the film.

In 1994, Sochay proposed an improved box office forecasting model based on regression analysis, and introduced R, the correlation coefficient square value, to detect the prediction effect. In the prediction analysis of the regression model, it is pointed out that the higher the coefficient is, the higher the accuracy of the prediction results is, and the smaller the coefficient is, the more deviating the box office prediction results are from the actual value. In 2006, Marshall proposed that advertising volume, number of copies of films, audience evaluation, number of theaters, audience age can be set as influencing factors. By using these factors, a multiple linear regression is explored to predict the number of tickets in Chilean film marke. Unfortunately, the model is only the number of viewers in different times of film release, rather than the prediction of the cumulative number of tickets in the whole screening cycle, which is not practical.

With the increasing prosperity of the film industry, the demand from the film industry promotes the development of box office research. Regression model is the main method of prediction and analysis. In addition to using the single traditional regression model, wenbin zhang and steven skiena collected 498 films released from 1960 to 2008 on internet movie database (imdb) data by combining the regression

analysis model with the k neighbor model (k-nearest neighbor model, k-nn). Through the quantitative analysis of film budget, screen number, box office in the first week, release period, mpaa rating, film production, type and whether or not sequel, this method combines imdb data and news data to predict the box office, which results in the best prediction effect. Meanwhile, the regression model has achieved good results in the prediction of low box office films, while the k-nn model has higher prediction accuracy for high box office films.

Due to the growth of the Internet, the number and types of video are more and more, the more abundant the data, the more challenges video prediction is facing. In 2010, Szabo made a breakthrough in the diversity of video numbers and types by using linear regression-based models to predict box office receipts by using Digg and YouTube broadcast data. As a key event showing box office prediction really entered the public field, google company published < quantifying movie magic with google search > in 2013. The film box office prediction method introduced in the film also uses multiple linear regression model. It is worth noting that in this paper, only using the single variable of film related search volume can explain the box office performance with an accuracy up to 70%. In order to further improve the prediction accuracy, the researchers added the characteristics of video search, advertising click rate and the number of films arranged in the hospital line in the model, and the predicted value coincided with the actual box office by 94%. The advantages of the model are simplicity, strong interpretation. However, too many regularization methods are used, causing the modeling easy to be overfitting.

In a word, regression model is a effective tool to predict popularity, to some extend. When analyzing a complex model, regression model is simple and convenient to implement. Besides, the correlation degree between each factor can be accurately measured and balanced. Thus it is easy to explain how data work during the process.

It should be noted that there are some limitations of this study since the training

dataset is too huge in practice. After all, the function of linear regression function ignores the interaction and nonlinear causality. If the feature dimension is too large and nonlinear, most of the prediction based on multivariate regression will be unreliable. A great deal of work has proved that the diversity of variables and the unpredictability of some variables limit the multivariate regression analysis, which results in the model overfitting.

2.1.3 Deep Learning

In recent years, deep learning has been widely used in industry, especially in the fields of image processing, speech recognition and natural language processing. Deep learning algorithm is a kind of multi-layer perceptual structure with multiple hidden layers. By combining low-level features to form an abstract representation in higher level, it is helpful to solve the problem of feature extraction. Powerful computing power, efficient parallelization framework and a large number of training dataset are the important foundation of achieving deep learning in practice.

CNN

At the 12-year imagenet conference, geoffrey hinton,ilya sutskever and alex krizhevsky from the University of Toronto submitted a deep convolution neural network algorithm called "alexnet". The error rate of graphic recognition is as low as 16%, more than 40% lower than that of the second place. It is the first time that CNN shows its super ability compared with traditional machine learning technology, such as SVM classifier.

Convolution neural network (CNN) is a kind of feedforward neural network, which has been successfully applied to the analysis of visual images in popular. The results show that individual cortical neurons only respond to stimulation in restricted areas called receptive fields. The sensory fields of different neurons can overlap and finally cover the whole receptive area. According to this theory, CNN can be built by

multi-layer neural network variants that requires minimal pretreatment. It is formed with three structural characteristics: local connection, weight sharing, and spatial or temporal sampling. These three characteristics make the convolution neural network have the ability of translation, scaling and distortion to a certain extent.

There are four main steps in CNN: convolution, subsampling (pooling) , activation and full connectedness. Following figure 2 shows the layout of a CNN structure.

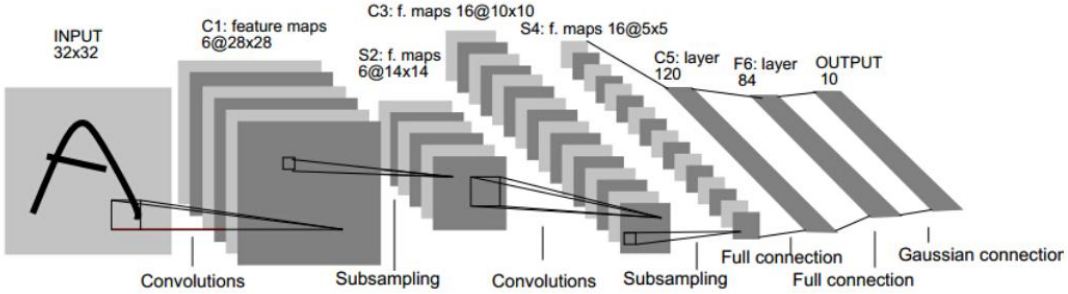

Figure 2 Layout of a CNN Structure

The first layers that receive an input signal are called convolution filters. Convolution is a process where the network tries to label the input signal by referring to what it has learned in the past. Due to the good property of being translational invariant,each convolution filter represents a feature of interest , and the CNN algorithm learns which features comprise the resulting reference. That confirm that the data feature do not disappear wherever the location is.

Subsampling is a step that can smooth the Inputs from the convolution layer in order to reduce the sensitivity of the filters to noise and variations. It can be achieved by taking averages or taking the maximum over a sample of the signal.
The activation layer controls how the signal flows from one layer to the next, emulating how neurons are fired in our brain. In order to improve and differentiate the efficiency of signal propagating, output signals which are strongly associated with past references would activate more neurons. CNN is compatible with a wide variety of complex activation functions to model signal propagation. For example, the most

common function is the Rectified Linear Unit (ReLU) which is favored for its faster training speed.

As for the last part in the network, there are some layers which are fully connected, meaning that neurons of preceding layers are connected to every neuron in subsequent layers. This mimics high level reasoning where all possible pathways from the input to output are considered.

RNN

Recurrent Neural Network (RNN) is a kind of neural network used to process sequence data. In the traditional feedforward neural network, the model can performance well if all inputs are independent of each other. But for many tasks, this is not a good solution. For example, a sentence has a complete grammatical structure and order, and each word in the sentence depends on the surrounding words. If a task requires the neural network is able to learn the meaning or sentiment in depth, the neural network must know the relevance between each word, especially the order of words.

Following figure 3 shows a layout of typical RNN structure. At each time *t*, the recurrent neural network gives an output for the input at that time combining with the state of the current model, and updates the state of the model. The input of A, the main structure of the recurrent neural network, comes from the input layer x_t, and there is a cyclic edge that provides h_{t-1} , the hidden state of the previous time . At each time, the module a of the recurrent neural network generates a new hidden state h_t after reading x_1 and h_{t-1} . The state h_{t-1} contains the information of the previous sequence from x_0 to x_{t-1}, which is used as a reference for the output o_t and then produces the output o_t at this current time. As the result of a property that the length of the sequence can be extended infinitely, it is impossible for the state h with limited dimension to save all the information of the sequence. Therefore, the model must learn to only retain the most important information related to the following tasks from

o_{t+1} to the end. Due to the design that the operation and variables in module A are the same at different times, recurrent neural network can be regarded as the result of infinite replication in the same neural network structure. The network is able to share parameters at different time positions, so that limited parameters can be used to deal with sequences of arbitrary time length.

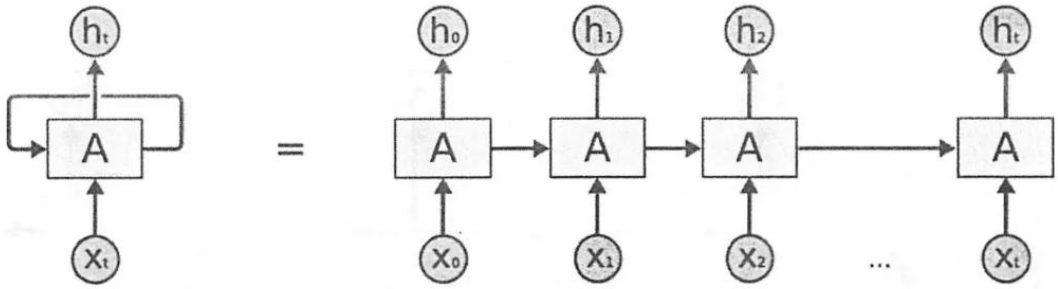

Figure 3 Layout of classical RNN model

However, as distance between words increases, RNN cannot use historical information efficiently. Fortunately, Hochreiter＆Schmidhuber proposed a special type of RNN called LSTM (Long Short-Term Memory) which is able to learn dependence between distant words. LSTM is carefully designed to avoid long dependency problems. Remembering longer historical information is actually their default behavior, not what they mainly study. The following figure 4 shows the layout of LSTM architecture. In brief, compared with single tanh function in a layer for RNN, there is a communication structure among four parts in a layer. Input gate and output gate control influence factors of different temporal memory. \The input gate with tanh function controls what information will be saved in the layer.The sigmoid function in forget gate determines what information will be thrown. At last, the output gate will generate the statement of this layer whose information has been filtered by a complex sigmoid and tanh function.

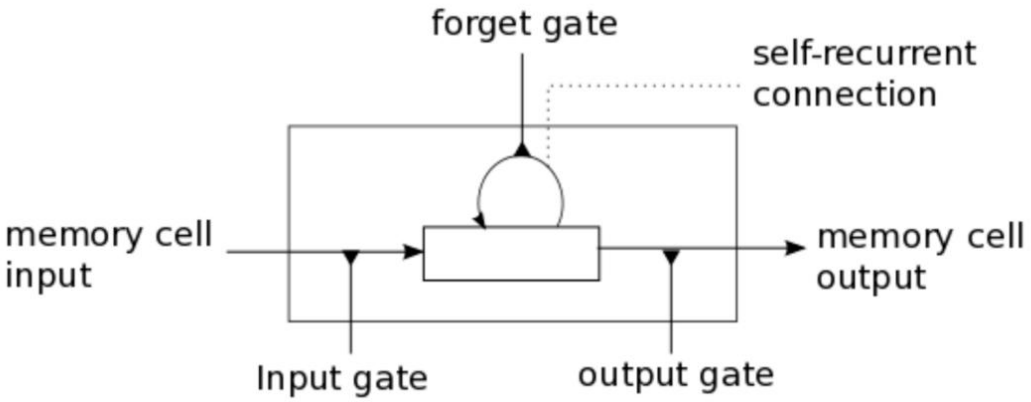

Figure 4 Layout of LSTM Architecture

2.1.4 Key Information Extraction

Key information extraction is an important research issue in opinion mining field. This can be traced back to the early time of literature retrieval. When full-text search was not supported, keywords can be used as words to search for this paper. Therefore, we can still see the key words in the paper. In addition, keywords can also play an important role in text clustering, classification, automatic summary and other fields. For example, when solving clustering classification issues, several documents with similar keywords can be regarded as a cluster, which can greatly improve the convergence speed of the clustering algorithm.

In general, there are two issues in keyword extraction field. The first issue is supervised learning algorithm. It treats the keyword extraction process as a binary classification problem. When a new document is processing, all the candidate words are extracted, and then the trained keyword extraction classifiers are used to classify each candidate words. The candidate words which are labeled as keywords are used as final output. The second issue is unsupervised learning algorithm. The candidate words are extracted and scored, and then the candidate words with the highest score are output as keywords. According to the scoring strategy, there are many practical

theories, such as tf-idf, textrank and other algorithms.

2.2 Tools and applications employed in project

2.2.1 Python

Python is a high-level programming language that combines interpretation, compilation, interactivity and object-oriented. Python is also very readable compared to other languages. Python was designed by Guido van Rossum in the early 1990s at the National Institute of Mathematics and computer Science in the Netherlands. So far Python 3.7 has been invented as the newest patch. In order to code Python, there are several convenient tools, such as PyCharm, Eclipse, which provide a development environment. As one of the most popular programming language, there are lots of useful code libraries that reduce the amount of programming work.

2.2.2 YouTube API

YouTube is an website that allows users to upload, watch and share videos. YouTube provides some great development tools called API, which allow developers access and analyze YouTube video data. Developers have to apply for a API key for individual research and pass the OAuth confirmation for data security. Then the API services can be used. For scraping YouTube video data, I used YouTube Data API v3 and YouTube Analytics API. Both of them can work well with Python.

Chapter 3 Problem Definition and Instrument Design

As mentioned above in 1.3.2 Project Objectives , the project explored an analysis instrument that is able to predict popularity of online videos and extract keywords based on comment text. As an Artificial intelligent issue, the project is required to research relevant theories, practice algorithms in experiment and finally give a solution.

3.1 Dataset

The issue requires a dataset that is huge enough for deep learning. By using YouTube API, I collected related information of over 2000 videos, containing view number and the comment text under each one. The comment text are sorted by agreed number. For the videos that containing more than 100 comments, only the most agreed 100 comments were download. These data are arranged as the original dataset. Later the dataset will be divided into training dataset, validation dataset and test dataset in training and evolution stages.

3.2 Prediction Tool Design

The instrument contains two modules. The first module is the prediction tool. This module consists of a text analysis part based on NLP technology and a regression Analysis part based on text parameters and view number.

For the NLP part, There is a model that has ability to solve a sentiment analysis task for comment text. In detail, firstly we need to process the dataset. By using libraries including *BeautifulSoup* and *textBlob*, the text data can be processed for removing punctuation, splitting into words and removing deactivation words. NLTK library is useful here to normalize the text. According to word embedding theory, I used word2vec tool to analyze these unsupervised data. Firstly, the domain emotion dictionary is constructed by using word2vec and word tags, and on this basis, negative words and degree adverbs are combined to calculate the sentiment tendency value of comments. Secondly, the comments with strong sentiment tendencies are selected as

the marked training set and the rest as the data sets to be classified. Finally, the machine learning method is used to generate classifiers for self-training until the end of the iteration. The training model is expected to give the probability that the comment sentiment is positive, which will be used as sentiment parameter in next steps. In general, there are many sentiment lexicons which have been created. However, these lexical resources are intrinsically non-contextualized, so it is necessary to improve their coverage based on given actual database. According to Milagros et al ‘s and Peter Turney’s paper, the analysis system is designed using the lexical. Finally, the system generate arrays of the sentiment parameters for each video.

As for the regression analysis part, I implemented several deep learning model to find the relationship between sentiment of comments and the view number. In previous sentiment analysis part, the data has been transformed as a group of sentiment parameters arrays, which is the input of this regression system. As the output, the view numbers are expected to be as closed as possible to actual number. However, the structure of data is so complex that it is hard to predict view number directly. Therefore, it is necessary to simplify the output form. I had to sort the view numbers and separate them from level 1 to level 10, representing the order of view numbers of each video from low to high. Then the prediction issue can be seen as a classification problem, which is tried to be an analogy with image recognition such as MNIST dataset recognition problem. LSTM, a kind of RNN structure, and classic CNN is implemented in this part. In order to reduce complexity, the LSTM is a forward LSTM structure, characterized by the fact that only the above information before the current word can be taken into account. The most reliable model will be tested and selected during experiment.

3.3 Keyword Extraction

As mentioned above, keyword extraction part is a supplement tool in the instrument system. The initial plan is to design an unsupervised model based on TextRank

algorithm. Due to time limitation, unfortunately, I have to renounce the idea yet. The keyword extraction system in the current instrument is designed based on simple tf-idf algorithm. In detail, all the comment texts under each video are combined, and then the five words with highest frequency are selected as output. Thus the experiment section will not include the part of keyword extraction because it is just a simple program.

At this point, the design of the instrument has been completed and implemented in general. In order to complete model training and compare the performances of all the model, some experiments are necessary which will be introduced in the following section.

Chapter 4 Experiments

Different approaches will be tested in order to improve the performance of this instrument as much as possible.

4.1 Sentiment Analysis Models Implementation Experiment

As introduced in above section, the sentiment analysis approaches in this project is related to analyzing lexicons. In the experiment, there are two different lexicons used in order to compare which one is better. These two lexicons are called sentiwordnet lexicon and afinn lexicon, which are both easy to implement in Python environment.

Each comment will be calculated and the model will generate 1, 0, or -1 corresponding positive, objective and negative attitude. There are 100 comments in each video set, so each video will be organized with a data that containing its id and its sentiment parameter array, which is the input of next regression model.

Due to unsupervised dataset, the performance of two models can not be compared so

far directly. The more reliable model will be determined in the following part.

4.2 Regression Models Implementation Experiment

After collecting the data from sentiment analysis model, we can begin to train a deep learning model to classify popularity of videos. As two of the most popular deep learning models, CNN and RNN are explored to achieve the goal of project.

It is a comparative Experiment of combined Model. According to different approach, The contrastive analysis is among four objects, including sentiwordnet + CNN, sentiwordnet + RNN, afinn + CNN, afinn + RNN. For this kind of classification experiment, accuracy is used as criterion. First of all, define and classify the correct classification function.

$$\delta(y_1, y_2) = \begin{cases} 1, & y_1 = y_2 \\ 0, & y_1 \neq y_2 \end{cases}$$

The correct number of categories in the classification results of all input data can be calculated according to the classification function .

$$accuracy = \frac{1}{N}\sum_{i=1}^{N} \delta\left(y_i, classify(\boldsymbol{x}_i)\right)$$

Where classify represents the classification method of the classifiers. N is the total number of input data. x i is the input data. y i is the category label of input data. According to the correct classification count, the accuracy of classification in all input data can be obtained.

Therefore, by comparing the accuracy of different models, we can select the best model combination for project. The result of experiment is shown in following figure 5.

Accuracy	sentiwordnet	afinn
CNN	73.69%	74.82%
RNN	67.37%	67.95%

Figure 5 Result of comparative experiment

As can be seen from the result, the combination afinn + CNN is the best experiment result. It reached 74.82% accuracy in prediction task. On the other hand, it can be found that the influence of sentiment lexicon is less than deep learning model. The difference of accuracy due to lexicon is only about 1 %. On the contrary, different deep learning algorithm can result in about 7% difference of accuracy.

Brief Summary

To sum up, the experiment explored the approaches of combining machine learning models for the prediction system in this project, and then analyzed the performances. In this chapter, the dataset is collected from YouTube comment and video information. The criterion is accuracy of prediction. The result shows that the weight of regression model is higher than sentiment lexicon. Besides, CNN is able to work better than RNN for regression task in the project.

Chapter 5 Legal and Commercial aspects

Legal Aspects

The final product must be only used for popularity and traffic value prediction. It must not exposure privacy of video publishers and video viewers. Any cited original algorithms and theories will be marked in report and reference list. Also any data created in the project will be carefully under YouTube Copyright.

Commercial Aspects

Online video industry is really a big and new market. There are millions of active users around world watching streaming media and unpredictable amount of potential video publishers. For example, YouTube is a popular international online video platform. A large number of video producers share their works on it. As of 2019, There are over 1.3 billion users on YouTube and in an average month, 8 out of 10

18-49 year-olds watch YouTube.

This project will be very useful for video publishers and platform managers. A few of the companies have developed parallel technology because of its high use value. It provides an effective tool to analyze feedback, and meanwhile, reduce manual analysis cost.

Chapter 6 Conclusion

6.1 Conclusions

In this document it is explored to invent an instrument that have ability to predict popularity and traffic value of online videos based on their comments. Overall, the author finished all the objective of this project. Following is the main work in summary:

(1) The paper practiced a sentiment analysis algorithm based on word embedding technology. Due to unsupervised data collected, word embedding method is thought a good approach to classify sentiment for short text. The method is characterized in that the probability of common occurrence of the same emotion words with the same polarity is large, and the probability difference of the same emotional words in the sentences of different emotional polarities is relatively large. The relevant emotional words in the field are identified, and the emotional words in the lexicon are expanded.

(2) The paper practiced a regression model based on deep learning technology. LSTM, a kind of RNN structure, and classic CNN was implemented to find the relationship between sentiment parameters and popularity of videos. Both of them have ability to classify labels according to input data. CNN focus on the absolute information of features contained in input data, while RNN tends to focus on the location of features in input data. Due to system structure, it is also an exploration

research involving the combination method of sentiment analysis model and regression model.

(3) A simple keyword extraction module based on tf-idf algorithm was also practiced in the paper. The module has ability to go through all the comment text of each video and extract five of the highest frequent words.

The final accuracy reached 74.82% which is, however, not a quite high accuracy. The main reason is that popularity of online videos is theoretically a multiple regression issue. Comment analysis is one of the most important and challenging input parameters in the multiple regression issue. In practice, a perfect popularity prediction tool have to consider about each aspects of the online video. The more characteristics are included, the more accuracy a prediction model might be. In conclusion, I think the project can be treated as a part of popularity prediction model for online videos. Moreover, the model tested in experiment is limited in my thought. More complex and individualized model may generate better results.

6.2 Further Work

First of all, this paper adopts the way of unsupervised learning for the sentiment analysis of comment text, which can be followed by the study of other methods with supervised learning. In theory, the effect of supervised learning methods will be better.

In addition, for the task of sentiment analysis of comment text, this paper chooses word embedding to participate in the machine learning framework, which is not comprehensive in NLP field. There are many other approaches, such as Naïve Bayes, Support Vector Machines and neural network. They should be considered a probability of the best solution for the system. Alternatively, the classification can be also replaced as the probability of positive attitude, then the sentiment parameters are represented as floating number, which may be more effective in regression model.

Furthermore, for the task of regression analysis, the experiments can be also extended. There are many improved neural network structures which may result in training a more accurate and reliable model. Meanwhile the neural network model also needs to change with input sentiment data. Due to time limit, I can not implement so many different methods in order to find a best solution for the system.

References

[1]Pozzi, F., Fersini, E., Liu, B., & Messina, E. (2016). Sentiment Analysis in Social Networks. San Francisco: Elsevier Science & Technology.

Zhang, L. , Wang, S. , & Liu, B. . (2018). Deep learning for sentiment analysis: a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1253.

[2]Trzcinski, T., & Rokita, P. (2017). Predicting Popularity of Online Videos Using Support Vector Regression. IEEE Transactions on Multimedia,19(11), 2561-2570.

[3]Liu Yan. (2012) Design and implementation of video playback prediction software based on sarima model. (doctoral dissertation, University of Electronic Science and Technology).

[4]Zhang, J., Zhang, Y., Ji, D., & Liu, M. (2019). Multi-task and multi-view training for end-to-end relation extraction. Neurocomputing, 364, 245–253. https://doi.org/10.1016/j.neucom.2019.06.087

[5]Bland J M, Altman D G. Statistic Notes: Regression towards the mean[J]. Bmj Clinical Research, 1994,
308(6942):1499-1499.

[6]Sochay S. Predicting the Performance of Motion Pictures[J]. Journal of Media Economics, 1994, 7(4):
1-20.

[7]Zhang W, Skiena S. Improving Movie Gross Prediction through News Analysis[J]. Web Intelligence,
2009, 01: 301-304.

[8]Szabo G, Huberman B A. Predicting the popularity of online content[J]. Communications of the ACM,
2010, 53(8): 80-88.

[9]Panaligan, R., & Chen, A. (2013). Quantifying Movie Magic with Google Search. Google Whitepaper: Industry Perspectives + User Insights, (June), 1–11.

[10] Osborne J W. Prediction in Multiple Regression[J]. Practical Assessment, 1999, 7(2):N/A.

[11]Davidian M. Hierarchical Linear Models: Applications and Data Analysis

Methods[J]. Journal of the
American Statistical Association, 1992, 88(463):767-768.

[12]Cai T, Hall P. Prediction in functional linear regression[J]. Annals of Statistics, 2006, 34(5): 2159-2179.

[13]Gonzalez, R. (2018). Deep Convolutional Neural Networks [Lecture Notes]. IEEE Signal Processing Magazine, 35(6), 79-87.

[14]Lecun, Y., Bottou, Bengio, & Haffner. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[15]Tian, Zhang, Li, Lin, & Yang. (2018). LSTM-based traffic flow prediction with missing data. Neurocomputing, 318, 297-305.

[16]Fan, M., Wu, G., & Jiang, L. (2012). Aspect Opinion Mining on Customer Reviews. In Proceedings of the 2011 International Conference on Informatics, Cybernetics, and Computer Engineering (ICCE2011) November 19–20, 2011, Melbourne, Australia: Volume 3: Computer Networks and Electronic Engineering (Vol. 112, Advances in Intelligent and Soft Computing, pp. 27-33). Berlin, Heidelberg: Springer Berlin Heidelberg.

[17]Mikolov, T., Joulin, A., Chopra, S., Mathieu, M., & Ranzato, M. (2015). Learning Longer Memory in Recurrent Neural Networks. ArXiv.org,ArXiv.org, Apr 16, 2015.

[18]Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & Javier González-Castaño, F. (2016). Unsupervised method for sentiment analysis in online texts. Expert Systems With Applications, 58, 57–75. https://doi.org/10.1016/j.eswa.2016.03.031

[19]Turney, P. D., & Littman, M. L. (2002). Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=edsarx&AN=edsarx.cs%2f0212012&site=eds-live