Speakers' Language Characteristics Analysis of Online Educational Videos

Dimitrios Kravvaris and Katia Lida Kermanidis

Department of Informatics, Ionian University, Corfu, Greece
jkravv@gmail.com, kerman@ionio.gr

Abstract. Research in the field of educational videos and the contribution of data mining to education can affect the instructors' approach to learning. This particular study focuses on online educational videos and more specifically on their speakers. Initially a survey is conducted related to the popularity of educational videos on the YouTube which are then divided into two categories the more popular and the less popular. Then the characteristics related to language are extracted from the transcript of the speakers and after a clustering procedure the differences between the two categories are stated. The characteristics related to the language of the speakers of the popular videos present very interesting results. That is, the pace of speaking is faster and the complexity off the sentences is higher than the ones in the less popular videos.

Keywords: Educational video, transcript, popularity, clustering, k-means.

1 Introduction

The advancement of social media adds a large amount of data on the web on a daily basis and especially in content-based communities such as YouTube and Daily Motion. A very large number of videos in social media concern education, and in many cases, constitute part of the traditional online courses [1] and the upcoming massive open online courses [2]. They are usually created by universities, companies, organizations or even individual users. In many cases transcripts of the video lectures are available.

The present research focuses on the study of educational videos from social media, oriented both to verbal content and to metadata of the pages that contain them. The present study thus, examines questions concerning issues such as why some educational videos are more popular than others and what are the characteristics that make a video popular. The issues arising are both interesting and complex. Our research innovation is that we examine them based on the audio language used in the educational videos. Through a qualitative study of the transcripts of the videos we extract the characteristics of the language used by the speakers (i.e. pace of speech, sentence length, commas, range of vocabulary etc.), which will be utilized in order to designate the speakers into to two basic types of speakers. These two types are based on whether the videos they take part in are popular or not. At this point another interesting question arises which we had to answer as well, i.e. what are these characteristics that define the popularity of a video after all, and how we can measure it? Using the

L. Iliadis et al. (Eds.): AIAI 2014, IFIP AICT 436, pp. 60–69, 2014.

metadata of the web pages that contain the videos we moved to an analysis related to the issue and propose a formula for defining video popularity. Finding the language characteristics of the speakers of the popular educational videos is very important both for the educational organizations and the individuals as creators of educational videos and for the scientific community since this study contributes to the research of linguistic data in the social media.

In the first part of the study we present relevant studies and point out how our research differs. In the second part we analyze the concept of popularity of online videos and we propose a formula for its estimation. In the third part we present the methodology used and include a thorough analysis of the characteristics used in our research. In the fourth part we present the experiments conducted as well as a commentary on the findings. Finally in the last part we present the findings of our research and how these can be utilized.

2 Related Work

As far as the videos are concerned, a lot of studies have been conducted in various field studies concerning video classification [3], [15] in order for the videos to fall automatically in certain categories using video and text data. Studies that concern the searching of videos and more specifically studies focusing on information retrieval browsing very large document collections [4] and video retrieval on the web utilizing the integration of multiple features [5], [7]. Finally, there have been studies that focus on video comparison [6] in order to estimate the percentage of visually similar frames.

The special characteristic of our research concerns the transcript of what the speakers say in each video. This has been used in other relevant studies concerning text mining such as text classification [13,14] and text clustering [8], [12], as well as studies concerning natural language processing [9,10]. Our research was inspired by the research conducted by Jin and Murakami [11] who studied the authors' characteristic writing styles as seen through their use of commas.

When it comes to social media, and more specifically YouTube, it has been shown that the introduction of videos in higher education has opened new horizons both to the educators who want to contribute to education and to learners who want to learn [16,17]. Thus, a new effort is being made in order for success in learning to be maximized. At this point our study comes in order to examine the educational videos in relation to their popularity on the YouTube. More specifically, we use the clustering method on metadata and on content data of the transcript of the video. Our purpose is to divide the videos in two categories: the most popular video category and the least popular videos category, and then to study which are qualitative speech characteristics of each category, regardless of the subject content of the videos.

3 Video Popularity

YouTube contains quite a few characteristics that could be utilized in order to define the popularity of a video, such as the number of views, of likes, of dislikes, the users' comments, the number of those who have chosen it as favorite and finally the number

of video responses [18,19]. The favorites and responses are the least used characteristics by the users. The views characteristic refers to the number of times the video has been viewed, while the likes, dislikes and comments can be used by registered users only. Especially for comments we should mention that we face two problems: the first concerns the complicated and time consuming procedure required in order to characterize the users' opinion [20,21], and the second concerns the ability to comment the video lecture, which could be deactivated by the creator and, thus, we would have no relevant comments.

Thus, in order to be fair concerning the videos in focus we chose to keep the characteristics that are definitely present and that attribute a positive value to the video. We ended up, therefore, using the views and the likes, in order to estimate the popularity of the videos. These two characteristics are based on human actions that show how many times a video has been viewed and how many people liked it.

We define as popularity P of a video i, which belongs to a certain category c, the normalized value of likes L and the number of views V according to formula

$$P_{i,c} = \frac{L_i}{maxL_c} + \frac{V_i}{maxV_c}$$

where $maxL$ and $maxV$ are the maximum values of likes and views correspondingly, that were observed in the particular video category. Since designating the value of P as *high* or *low* is subjective, we used the mathematical method of median [31]. The median is a measure of central tendency. In our case it represents the value for which half of videos' popularity are higher and the other half are lower. In that way splitting in half the videos of high and low popularity we can use machine learning methods in order to extract knowledge concerning what makes a video more popular than another.

We chose not to use the lifetime of a video on YouTube as a parameter in estimating the popularity of a video, because there seemed to be a problem: The new videos with few likes and views seemed to be more popular which was wrong because older videos had more likes and views.

4 Methodology

4.1 Data

Our data were collected from YouTube, which is the third most visited social media site worldwide [22] and the largest provider of videos [23]. YouTube provides its users with a specific space to upload videos that fall into the category of educational videos[1]. Searching through the category of Education of YouTube by inserting keywords from different scientific fields such as computer science, physics, medicine, art, health, philosophy, energy and others, 20830 videos were collected among which 1108 (5.3%) had English transcripts. The total duration of the 1108 videos used in our research is 473 hours and have over 242 million views in total. From each video metadata attributes were collected using the YouTube API v2 [24] as well as qualitative

[1] `http://www.youtube.com/education`

attributes of speech (after processing the transcripts of videos). Grzybek's et al. [32] and Mahowald's et al. [33] research shows the importance of words, Hill's and Murray's research [34] note the value of commas and Palmer's study [35] highlights the importance of sentence segmentation of a natural language text. Thus, we used these important structural elements for the definition of our qualitative attributes. Below we refer to these attributes and their description in categories.

- Metadata
  In this category there are two attributes. The first one is the *Duration* attribute which refers to the second of the total appearance of the online video. The second attribute is the *AuthorUri* which concerns the unique identity of the owner of the video on YouTube, which may refer to a University, an educational organization or an individual. Both attributes come from metadata of the YouTube page, which contain the video in focus.
- Words
  The words category contains the qualitative characteristics of the transcript of the educational video. More analytically, the attribute *NumOfWords* concerns the number of words used by the speakers of the video. This attribute shows the real duration of speech, since we count neither the duration of speech, which contains times pauses, nor the duration of a video which contains other elements such as ads or short introductions before the educational video begins. The second attribute *AvgWordLength* concerns the average word length. This attribute helps us form a complete view of the net length of speech we referred to earlier, since videos differ also in the length of words used, besides the number of words.
- Transcript Sentences
  This category contains four attributes which concern: the number of the transcript sentences (*NumOfSent*), the minimum sentence length in characters (*MinSentLength*), the average sentence length in characters (*AvgSentLength*) and the maximum sentence length in characters (*MaxSentLength*). All these four attributes describe the number and the length of the transcript sentences. Thus, through the transcripts we can extract qualitative information concerning the sentence length used by the speakers, supposing that longer sentences are more likely to contain more information for the listener that shorter ones.
- Sentences complexity
  This category contains two attributes concerning the commas contained in the transcripts. The *NumOfCommas* attribute refers to the total number of commas contained in the transcript while the *AvgNumOfCommasPerSent* attribute shows the average number of commas per sentence. Commas are used in order to avoid ambiguity. They are mainly used in lists, for separation causes, to set off certain adverbs at the beginning of a sentence and in parenthetical phrases. All the above indicate that a sentence with commas is more complicated in structure and in meaning that one without commas [30].
- Vocabulary
  This category contains the *NumOfUniqueWords* attribute which shows the number of unique words in the transcript. The more unique words a transcript contains the wider the vocabulary used by the speaker, without it necessarily being more advanced since the videos come from different scientific fields and contain the domain-specific terminology of the corresponding fields.

- Flow of words
  This category contains two attributes: the *MicroRhythm* attribute that refers to the micro flow of words and the *MacroRhythm* attribute which refers to the macro flow of words. More analytically, the *MicroRhythm* attribute measures the average flow of words in the time (measured in seconds) the corresponding transcript text is displayed on the screen, and the *MacroRhythm* attribute measures the flow of words in the total time the transcript texts are displayed on the screen.
- Evaluation
  This category contains the evaluation attribute of our study named *Popularity*. This attribute is used for the binary classification of the educational videos. It has two values *high* and *low* as it was described in the previous section of the present paper.

4.2 Experimental Procedure

In the beginning we conduct a statistical analysis of our data. At this point we study any extreme cases and we suggest solutions to deal with them. The purpose is to pre-process our data so as to avoid problems during the experimental procedure, such as missing values in the data of our datasets.

In the experimental procedure we used the Weka version 3.6.10 software [25]. We employed unsupervised learning methods for the clustering experiments. More specifically, a centroid-based clustering algorithm [26] using SimpleKMeans, with the Euclidian distance function [27] has been used. The SimpleKMeans method is quite suitable for our experiments since it is easy to understand and to explain its clustering outcome [28]. Two clusters were chosen for the value of K (in SimpleKmeans), since we have two class values: high/low popularity. Moreover, we chose to use the clustering mode classes-to-clusters evaluation [29], which assigns classes to the clusters based on majority and computes the classification error of the videos that have different value from the class value of the cluster they belong to. With the above procedure, on the one hand, we can study the differences between the qualitative characteristics of the videos (that come from the transcript), and, on the other hand, to evaluate how these characteristics can define the videos' popularity.

5 Experimental Results

5.1 Data Analysis

While analyzing our data we found out that there is a great difference in the duration of videos and for this reason we have discretized their duration in 10-minute intervals. The results are presented in figure 1 below, which shows the number of videos in each time category they belong to. We find that the greatest number of educational videos fall into the 1 to 10 minute category (47.5% of videos), while the 41 to 50 minute and 51 to 60 minute categories contain 25% of videos in total. In the first case, there are short videos concerning the time duration, while in the second case long ones, for this reason, thus, we divided the initial dataset into two new ones based on their duration. In this way we can conduct our study on data that have similar characteristics.

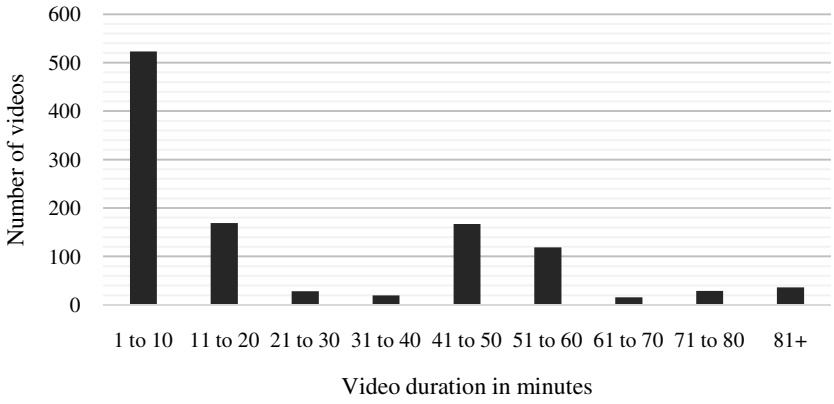

Fig. 1. Distribution of videos based on time duration

5.2 Short Videos

After conducting our experiment on the dataset that contains short (in terms of duration) videos we extract the following results shown on table 1. The videos that belong to *Cluster-0* are in majority of high popularity, while the videos that belong to *Cluster-1* are of low popularity. It should be mentioned that 66.54% of the videos have been correctly clustered. Considering that we have to do with data that are based on the human activity of speech the percentage can be characterized as highly positive.

Table 1. Clustering results for short videos

Attribute	Cluster-0 (209 videos)	Cluster-1 (317 videos)
Duration	398,0909	179,7476
AuthorUri	TEDEducation	Udacity
MicroRhythm	3,2269	3,0254
MacroRhythm	2,9277	2,6793
NumOfWords	1064,9809	417,3817
AvgWordLength	4,5849	4,6857
NumOfSent	64,8852	27,1735
MaxSentLength	479,3923	323,6498
MinSentLength	19,3541	35,2114
AvgSentLength	141,7913	124,3602
NumOfCommas	56,3636	19,4795
AvgNumOfCommasPerSent	0,9938	0,7671
NumOfUniqueWords	383,6268	200,6467

Based on the qualitative characteristics of the transcripts of the video speakers, as they were described above, we can describe the types of the speakers of each cluster. More specifically, the speakers of the popular videos (*Cluster-0*) present the following language characteristics, compared to speakers of the less popular videos (*Cluster-1*).

- Greater net length of speech (*NumOfWords, AvgWordLength*).
- Sentences with more information for the listener (*NumOfSent, MaxSentLength, MinSentLength, AvgSentLength*).
- More complex sentences (*NumOfCommas, AvgNumOfCommasPerSent*).
- Greater number of unique words (*NumOfUniqueWords*).
- Faster pace of flow of speech (*MicroRhythm, MacroRhythm*).

Thus, on the case of short videos, in order for the educational video to be popular among internet users, the speaker should speak at a fast pace and use long and complex sentences, so as to take greater advantage of the time he is given to inform the audience about the issue in focus. The fact that users prefer to listen to a fast pace speaker is very interesting which means that they are closely paying attention to what the speaker is talking about and they are fully focused on the subject of interest in order to follow the speaker's speech pace.

5.3 Long Videos

From the experiment conducted on the dataset that contains long (in terms of duration) videos we extract the following results shown on table 2. In this case, also, as was also shown in the previous experiment, videos that belong to *Cluster-0* are in majority of high popularity, while videos that belong to *Cluster-1* are of low popularity. The correctly clustered videos reach 86.36%, which is extremely positive for the classification of the videos.

Table 2. Clustering results for long videos.

Attribute	Cluster-0 (167 videos)	Cluster-1 (119 videos)
Duration	2965,9641	2974,3529
AuthorUri	MIT	YaleCourses
MicroRhythm	3,0471	2,6756
MacroRhythm	2,3240	2,4421
NumOfWords	6779,0719	6684,5630
AvgWordLength	4,2704	4,6612
NumOfSent	465,0659	364,9244
MaxSentLength	392,9521	818,7227
MinSentLength	1,9162	10,3193
AvgSentLength	77,5774	130,3118
NumOfCommas	461,1617	322,4202
AvgNumOfCommasPerSent	1,0453	0,9979
NumOfUniqueWords	974,8323	1343,6555

Following the same logic, as in the previous case of short videos, we can describe the types of speakers of every cluster. In this way, we can record comparatively the language characteristics of the speakers of the popular videos (*Cluster-0*) compared to the speakers of the less popular videos (*Cluster-1*). We find out that the speakers of the popular videos have:

- Practical the same length of speech as the speakers of the less popular videos. (*NumOfWords, AvgWordLength*).
- Sentences containing less information for the listener (*NumOfSent, MaxSentLength, MinSentLength, AvgSentLength*).
- More complex sentences (*NumOfCommas, AvgNumOfCommasPerSent*).
- Lower number of unique words (*NumOfUniqueWords*).
- Faster pace of micro flow of words (*MicroRhythm*), and almost the same pace of macro flow of words (*MacroRhythm*) as the speakers of less popular videos. This means that on average the speaker in a popular video uses more words at a given period of time.

To sum up, in order for a long video to be frequently viewed and positively reviewed, the speaker has to speak at a fast pace, to limit his vocabulary to the issue in question and to use complex sentences, which, however, do not carry too much information. In that way the user stays focused on the speaker's words and does not get confused or bored while watching the video.

5.4 Similarities

Based on our findings, there are similarities between the characteristics of the speakers of the popular short videos and the popular long videos as these are shown in column *Cluster-0* of tables 1 and 2. The similarities concern: a) the pace of speech, where we see that the speakers use almost the same number of words at a given period of time and b) the complexity of the sentences, where it is shown that speakers prefer to use more complicated in structure and in meaning sentences for their listeners. Thus, we conclude that, in order to create a popular educational video, regardless of its duration, the main speaker has to have a good command of the audio language and to be fully aware of the lecture subject so as to be able to express complex issues at a fast pace of speech. Knowing that the level of knowledge of the English language of the video listeners varies, the characteristics mentioned above seem to be very important in order for the lecture subject to be effective.

6 Conclusion

In our research we studied the language characteristics that a speaker of an educational video should have in order for the video to be more acceptable by the users of the social media. The whole procedure was based, on the one hand, on the qualitative research of the video transcripts, from which the language characteristics were extracted, and, on the other hand, on the classification of the videos in categories according to their popularity.

The popularity of the videos constituted the first part of an interesting analysis for our research. The formula suggested was based on the likes and views attributes, which besides YouTube, appear in other social media. The classification of the videos in most popular ones and least popular ones, based on the median method, ensured that this classification was objective.

Through our experimental procedure one can find out that in short videos speakers use speech more effectively. Speaking at a faster pace and using sentences with more

information do not allow time to be wasted and help keep their listeners' interest. On the other hand, in popular long videos the speech contains more complex sentences than the speakers in popular videos. Finally, we show that both type of speakers in long and short popular videos have similar fast pace of speech. Their common characteristic that concerns the fast pace of speech seems to be interesting and urges us to study further what the pace should be in order for a listener to be satisfied.

The education industry uses videos as a basic tool. Our research shows that new knowledge can be extracted through machine learning techniques from the large quantity of free data in social media. In this way it is possible to identify the factors that can conduce to creating more popular and higher quality educational videos.

References

1. Moel de, E.L.: Expanding the usability of recorded lectures (2010), http://purl.utwente.nl/essays/59431
2. Waldrop, M.M., Nature Magazine: Massive open online courses, aka MOOCs, transform higher education and science (2014)
3. Gibbon, D.C., Liu, Z.: Introduction to Video Search Engines. Springer (2008)
4. Cutting, D.R., Pedersen, J.O., Karger, D.R., Tukey, J.W.: Scatter/gather: a cluster-based approach to browsing large document collections. In: SIGIR, pp. 318–329 (1992)
5. Yang, J., Li, Q., Wenyin, L., Zhuang, Y.: Searching for flash movies on the web: A content and context based framework. World Wide Web 8(4), 495–517 (2005)
6. Cheung, S.C.S., Zakhor, A.: Efficient video similarity measurement with video signature. IEEE Trans. Circuits Syst. Video Technol. 13(1), 59–74 (2003)
7. Hindle, A., Shao, J., Lin, D., Lu, J., Zhang, R.: Clustering web video search results based on integration of multiple features. World Wide Web 14(1), 53–73 (2011)
8. Amine, A., Elberrichi, Z., Simonet, M.: Evaluation of text clustering methods using wordnet. Int. Arab J. Inf. Technol. 7(4), 349–357 (2010)
9. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. The Journal of Machine Learning Research 12, 2493–2537 (2011)
10. Friederici, A.D.: The brain basis of language processing: from structure to function. Physiological Reviews 91(4), 1357–1392 (2011)
11. Jin, M., Murakami, M.: Authors' characteristic writing styles as seen through their use of commas. Behaviormetrika 20(1), 3–76 (1992)
12. Shehata, S., Karray, F., Kamel, M.S.: An efficient concept-based mining model for enhancing text clustering. IEEE Transactions on Knowledge and Data Engineering 22(10), 1360–1371 (2010)
13. Passini, C., Luiza, M., Estébanez, K.B., Figueredo, G.P., Ebecken, F., Nelson, F.: A Strategy for Training Set Selection in Text Classification Problems. International Journal of Advanced Computer Science & Applications 4(6) (2013)
14. Kiritchenko, S., Matwin, S.: Email classification with co-training. In: Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research, pp. 301–312 (2011)
15. Filippova, K., Hall, K.B.: Improved video categorization from text metadata and user comments. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 835–842 (2011)

16. Gilroy, M.: Higher education migrates to YouTube and social networks. Education Digest 75(7), 18–22 (2010)
17. Selwyn, N.: Social media in higher education. The Europa World of Learning (2012)
18. Chatzopoulou, G., Sheng, C., Faloutsos, M.: A first step towards understanding popularity in youtube. In: INFOCOM IEEE Conference on Computer Communications Workshops, pp. 1–6 (2010)
19. Figueiredo, F., Almeida, J.M., Gonçalves, M.A., Benevenuto, F.: On the Dynamics of Social Media Popularity: A YouTube Case Study. arXiv preprint arXiv:1402.1777 (2014)
20. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies 5(1), 1–167 (2012)
21. Chen, H., Zimbra, D.: AI and opinion mining. IEEE Intelligent Systems 25(3), 74–80 (2010)
22. Wattenhofer, M., Wattenhofer, R., Zhu, Z.: The YouTube Social Network. In: ICWSM (2012)
23. Moran, M., Seaman, J., Tinti-Kane, H.: Teaching, Learning, and Sharing: How Today's Higher Education Faculty Use Social Media. Babson Survey Research Group (2011)
24. Padilla, A., DeFields, A.: Beginning Zend Framework. Apress (2009)
25. Weka.: Data Mining Software in Java. University of Waikato (2014), http://www.cs.waikato.ac.nz/ml/weka/
26. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C., Silverman, R., Wu, A.Y.: The analysis of a simple k-means clustering algorithm. In: Proceedings of the Sixteenth Annual Symposium on Computational Geometry, pp. 100–109 (2000)
27. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 881–892 (2002)
28. Vora, P., Oza, B.: A Survey on K-mean Clustering and Particle Swarm Optimization. International Journal of Science and Modern Engineering (IJISME), 24-26 (2013)
29. Färber, I., Günnemann, S., Kriegel, H.P., Kröger, P., Müller, E., Schubert, E., et al.: On using class-labels in evaluation of clusterings. In: MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD (2010)
30. Wiegand, N.: Creating complex sentence structure. In: Proceedings of the Annual Meeting of the Berkeley Linguistics Society vol. 10 (2011)
31. Beliakov, G., Bustince, H., Fernandez, J.: The median and its extensions. Fuzzy Sets and Systems 175(1), 36–47 (2011)
32. Grzybek, P., Stadlober, E., Kelih, E.: The relationship of word length and sentence length: the inter-textual perspective. In: Advances in Data Analysis, pp. 611–618. Springer (2007)
33. Mahowald, K., Fedorenko, E., Piantadosi, S.T., Gibson, E.: Info/information theory: Speakers choose shorter words in predictive contexts. Cognition 126(2), 313–318 (2013)
34. Hill, R.L., Murray, W.S.: Commas and spaces: The point of punctuation. In: 11th Annual CUNY Conference on Human Sentence Processing (1998)
35. Palmer, D.D.: Tokenisation and sentence segmentation, chapter 2. In: Dale, R., Moisi, H., Somers, H. (eds.) Handbook of Natural Language Processing. Marcel Dekker (2000)