Peem Lerdputtipongporn

36757: Advance Data Analysis I

Last updated: 6th April 2022

**Statistics advisor(s):** Nynke Niezink and Larry Wasserman (Department of Statistics and Data Science, CMU).

**External scientists names and affiliations:** Rema Padman (Heinz College of Information Systems and Public Policy, CMU); Anjana Susarla (Broad College of Business, Michigan State University); Xiao Liu (WP Carey Business School, Arizona State University).

**Section 1 – Introduction**

**Introduction**

During the past decade, annual healthcare spending in the US has grown exponentially. The spending reaches $4.1 trillion — 19.7% of the United States’ GDP in 2019 — and is projected to reach $6.2 trillion by 2028 (CMS). One of the major factors contributing to the rising medical expenditure is the lack of health literacy required to understand medical information, most of which is presented in text. Nowadays, Youtube offers alternative means for physicians to communicate with patients through audios and graphics. Moreover, it serves as a platform where patients can produce content as well as interact with one another. However, the rise of Youtube as alternative means to promote health literacy and disseminate medical information poses a fundamental question: how should one identify if a health video contains misinformation? Answering this question allows healthcare providers to better assist patients by directing them toward clinically appropriate treatment. In this paper, we focus specifically on classifying whether a Youtube video about diabetes contains misinformation.

Diabetes mellitus, also known as diabetes, is a chronic, non-communicable disease whereby defects in insulin discretion cause prolonged high blood glucose (Kharroubi and Darwish 2015). 9.1% of Americans aged 20 years and above suffer from diabetes, and people with diabetes lose their life expectancy by 0.89 years (Preston et. al. 2018). Some scholars have studied how misinformation about acute diseases such as COVID-19 spread through social media [citation]. However, there is relatively little literature on misinformation about chronic diseases. The scarcity of such literature can be attributed to the fact that assessing the quality of health videos is time-consuming and requires people with medical expertise. To circumvent this problem, Liu et al. (2020) have proposed using semi-supervised learning to classify video in term of its understandability and medical content as well as identify how both qualities affect viewers’ engagement (2019, 2020). Using each video’s features and metadata to evaluate its quality, their method scales to large dataset.

However, a video is understandable and contains high medical content does not imply it should be recommended to patients. In particular, if the video contains inaccurate medical information, it may prevent patients from seeking appropriate treatment. While physicians may evaluate misinformation in health videos using some guidelines, they cannot manually evaluate the accuracy of every video. Moreover, from a content-creator perspective, it is helpful to know what makes their videos medically inaccurate and how to improve the quality of their videos.

In other words, a better understanding of the relationship between video features/metadata and its degree of misinformation is needed, which we will explore in this paper.

Our work extends the framework by Liu et. al. by using semi-supervised learning to impute labels on misinformation. To do so, we [the external scientists] retrieved corpus of words related to diabetes from United States Public Health Service (USPHS), used them as keywords to search diabetes videos on Youtube, collected the metadata of ach video (such as number of likes, comments, viewCounts, likeCounts, etc.), and used Google Intelligence/BLSTM to extract medical-related terms. Our analysis shows that [to be inserted], and its implication [not yet determined] will be discussed.

**Literature Review**

Many scholars have attempted to study online text misinformation (Muric et. al.’21, Tang et. al. ‘21). Misinformation spreading through text often involves acute diseases, such as seasonal flu, COVID-19, among others. Other scholars attempt to characterize misinformation through user-generated labels on online platform (Knuutila et. al. ’20) or infer misinformation based on the medium network structure (Tang et. al. ’21). Papadamou has written a comprehensive review of current literature involving abhorrent content or misinformation on Youtube (2021). To the best of the author knowledge, we have identified one paper studying Youtube misinformation (Jagtap et. al. ’21). Jagtap et. al. claimed video metadata are not indicative whether a video contains misinformation or not, proposing a natural language processing approach and synthetic minority oversampling technique to analyze the video content instead. Liu et. al. have studied the effect of video understandability and degree of information on user engagement (2019).

**Section 3 -- Results**

**Subsection 3.1**: Description of the data, EDA plot/table/etc + comments. Do not include plots/tables/etc that you write no comments about (if you have nothing to say, don't show it. If you show it, walk us through it).

To curate this dataset, we (Rema Padman) query Youtube under the incognito mode using keywords related to diabetes. The Top 50 search results are stored in our dataset. Our dataset consists of 11,483 videos, of which 573 are labelled according to PEMAT guideline on understandability, actionability, and whether they contain medical information. We do not have the misinformation labels yet. The distribution of each video falls under the following groups:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Understandable (1)** | **Not understandable (0)** | **Total** |
| **Contains medical information (1)** | 348 | 39 | 387 |
| **No medical information (0)** | 84 | 102 | 186 |
| **Total** | 432 | 141 | 573 |

Aside from the labels, our dataset contains 50 features (to be checked due to redundant labels). These features can be grouped into the following categories, to name a few:

|  |  |  |
| --- | --- | --- |
| **Video Metadata** | **Video Content** | **Derivatives of video metadata** |
| Caption ID  Number of likes, dislikes, views.  Duration.  Date of publishing.  Channel information. | Category.  Video topic  Subtitle  Title  Speech confidence  Scene count | Cosine similiarity between keywords, titles, and description.  Description readability. |

**Current findings**

1. The number of word counts in video description has little to no correlation with whether the video is understandable, actionable, or contains medical information (see Figure 3.1)

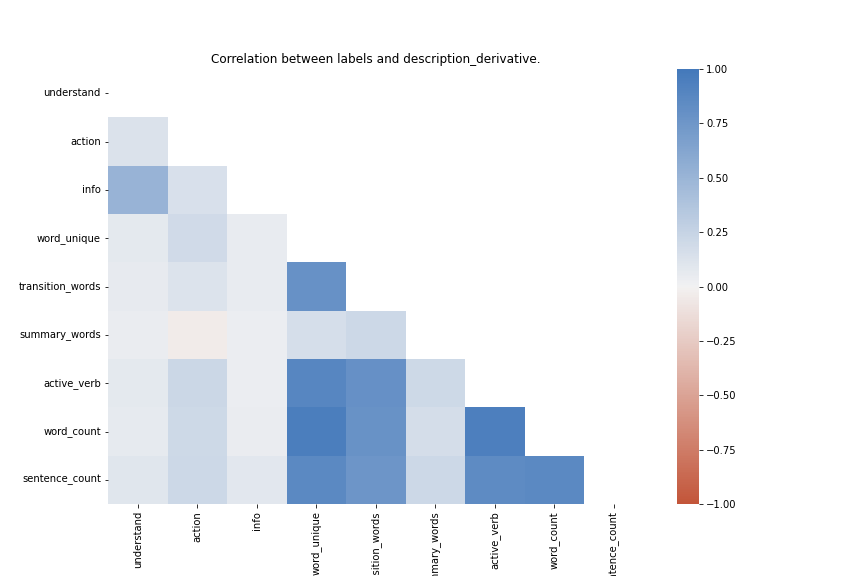


Figure 3.1: A heat map (correlation plot) between various word counts

and the video labels. The light blue region in left side suggests weak correlation.

1. There is little difference in feature value distribution between understandable vs not-understandable, actionable vs not-actionable, and informative vs not-informative video.

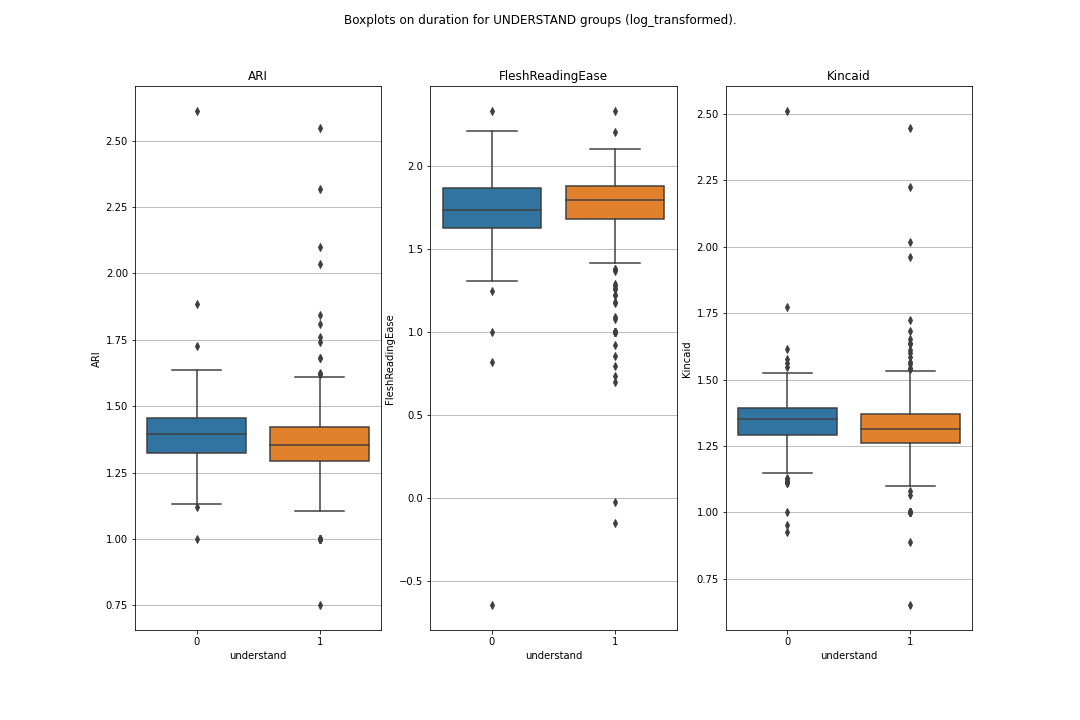


Figure 3.2: A box plot comparing distributions of readability indices (log-transformed) between understandable and not-understandable videos.

1. Our dataset is extremely imbalanced. 75% of the videos are understandable; 68% actionable; 50% informative. Therefore, fitting unweighted off-the-shelf classifier will lead to extremely low detection of videos falling in Class 0.

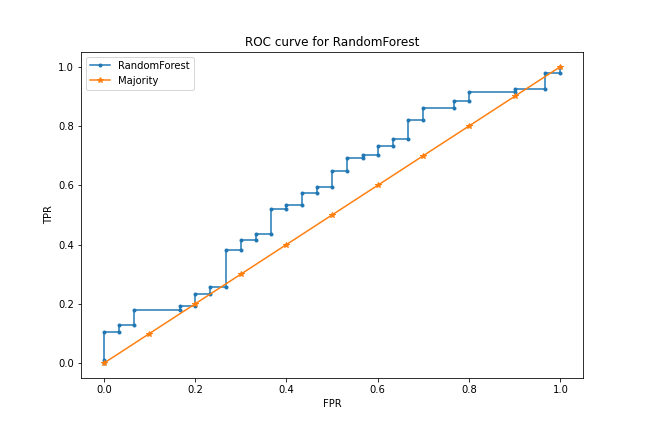


Figure 3.3: An ROC curve of fitting randomForest to classify video understandability. The blue ROC curve suggests that the classifier, trained

on unweighted dataset, has little prediction capability.

1. Our dataset consists of 1,025 videos unrelated to diabetes. These videos appear under queries that are not language-specific, such as nesina (a diabetes drug; also the name of a famous rapper).

**Section 4: Conclusions**

TBD

**Appendix 1:** Description of all variables in the dataset.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Group** | **Description** |
| captid (string) | Metadata | ID of Youtube caption. |
| captsLastUpdated (text, number): | Metadata | The last time a video caption is updated; is stored in the YYYY-MM-DD: Time format. |
| categoryId (encoding of text) | Content | The category each Youtube video belongs to, such as person, blog, science. See Youtube API for more info. <https://gist.github.com/dgp/1b24bf2961521bd75d6c> |
| contentDefinition (string) | Content | Consists of two types: hd (high definition) or sd (standard definition). 68% are hd. |
| 'contentLicensed (Boolean) | Content | Whether the content on that channel is licensed. If a video is licensed, it cannot be used for commercial purposes without the permission of the video creator. |
| description (string) | Metadata | Description of videos seen below the rectangle box (e.g., importance of lipid metabolism). |
| keyword (text) | Search | Keyword used to query that video. Some videos appear under more than one keyword. |
| 'publishedAt' (text, number): | Metadata | Date and time a video is published. |
| rank (number) | Search | The rank at which the video appears if searched using keywords in incognito mode. |
| relevantTopicIds (encoding of text): | Content | TopicIDs created by Youtube based on knowledge graph. |
| 'subtitle' (text): | Content | The transcript of what is said in the video (e.g., “In this video, we will talk about”). In this dataset, only 700 videos have subtitles. |
| 'title' (text): | Metadata | Title of the video (e.g., “Importance of insulin”). |
| 'video\_duration (text, number) | Metadata | Duration of the video |
| 'topicIds' (encoding of text): | Content | All except 3 are NaN |
| 'Kincaid' (number)\*\*: | Derivative | Flesch-Kincaid readability index. Measures the minimum education level required to comprehend the text.  \*\* The higher Kincaid is, the more difficult a text is to comprehend. See formula using link above. |
| 'FleshReadingEase (number)\*\*: | Derivative | Flesch-reading ease index. Measures how easy it is to read a text. The higher the reading score, the easier a piece of text is to read) |
| 'ARI' (number)\*\*: | Derivative | Automated readability index. The higher ARI is, the more difficult it is to read the text. |
| 'word\_count' (number) \*\* | Derivative | How many words appear in the description; calculated using Python library.  \*\* NOTE: Cannot handle special symbols, related to readability. |
| 'word\_unique (number)\*\* | Derivative | Number of unique words in the description of the video. |
| 'transition\_words (number)\*\*: | Derivative | Number of transition words (think: and, so, but, however, etc.) in the description of the video. |
| 'summary\_words (number)\*\* | Derivative | Number of transition words (think: and, so, but, however, etc.) in the description of the video. |
| 'active\_verb' (number)\*\* | Derivative | How many active verbs appear in the description. |
| 'sentence\_count' (number) \*\* | Derivative | How many sentences appear in the description. This is calculated using Python library. |
| channelCommentCount (number) | Metadata | How many comments a video receives.  Some channels have extremely high commentCounts, such as British GOT. |
| channelDescription (text): | Metadata | Description of each channel (created by channel owner in About). |
| channelId (text) | Metadata | ID of the video’s channel. |
| channelPublishedat (text): | Metadata | Time at which channel was established. |
| channelSubscriberCount (number): | Metadata | How many subscribers a channel has. |
| channelTitle (text) | Metadata | Name of the channel (e.g., All about Diabetes and Related). |
| channelVideoCount (number): | Metadata | How many videos that video’s channel has posted. |
| channelViewCount (number): | Metadata | How many views a channel posting that particular video has received. |
| 'keyword\_title\_cosine' (Number) | Derivative | Cosine similarity between title and keyword. Doesn’t suffer from out of range problem. |
| 'keyword\_decription\_cosine' (Number): | Derivative | Cosine simiarlity between keyword and description. |
| comment\_title\_cosine' | Derivative | Cosine simiarlity between comment and title. |
| Comment\_description\_  cosine' (Number) | Derivative | Cosine simiarlity between comment and description. |
| viewCount (number) | Metadata | How many views the video receives. |
| likeCount (number): | Metadata | How many likes a video receives. |
| dislikeCount (number) | Metadata | How many dislikes a video receives. |
| commentCount (text) | Metadata | How many comments have been made on each video. |
| postive\_comment\_count' (Number) | Derivative | Of the most recent 100 comments, how many are positive comments? |
| Negative\_comment\_  count' (Number) | Derivative | Of the most recent 100 comments, how many are negative comments? |
| neutral\_comment\_count' (Number): | Derivative | Of the most recent 100 comments, how many are neutral comments? |
| comment\_unique\_words(Number) | Derivative | The number of unique words appearing in the most recent 100 comments. |
| comment\_total\_words' (Number): | Derivative | The total number of words appearing in the most recent 100 comments. |
| Scene\_count (Number) | Derivative/Content | The number of scenes appearing throughout the video. |
| Object\_count (Number) | Derivative/Content | The number of medical objects appearing throughout the video. |
| Text\_confidence (Number) | Derivative/Content | The average confidence level in Youtube auto-generated transcription. |
| Speech\_confidence (Number) | Derivative/Content | The quality of audio level in Youtube subtitle. |

**Citations:**

<https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet>

Holman: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7077778/>