Peem Lerdputtipongporn

36757: Advance Data Analysis I

Last updated: 15th February 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).

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/>

**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.

>> (To Nynke and Larry) Question about this paragraph: Should I describe Liu et. al. more in details? I believe what I wrote suffices at a high-level. However, since we are building off of their work, it feels odd to summarize their work in two sentences. Should I delve into the weeds of mapping features to PEMA guideline, specific algorithm, etc.?

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) [TODO: Check], used them as keywords to search diabetes videos on Youtube, collected the metadata of each video (such as number of likes, comments, viewCounts, likeCounts, etc.), and used Google Intelligence/BLSTM to extract medical-related terms. [TODO: Describe methods] Our analysis shows that [to be inserted], and its implication [not yet determined] will be discussed.