A Trade-off Between Accuracy and Exposure:

What Does Formality Mean in the Context of Science Communication on YouTube?

A transcript analysis of the science communication genre on YouTube.

Master Thesis – MSc. Business Information Management (Data Science Track)
Erasmus University Rotterdam | Rotterdam School of Management

Coach: Dr. D.K. Zegners | 2nd Reader: M. Ansarin MSc
Gerbrand van Dijk (539179gd) – Date: Friday, 23rd July 2021

ABSTRACT

As citizens in a democratic society, being well-informed allows us to live more free and independent lives through the accurate discernment between the true and false. Allowing citizens to be well-informed is exactly what the main premise of science communication is. Hence, optimizing its delivery is something with which there is much we stand to win. In this study we hypothesize that altering how formally scientific information is delivered, can help optimizing its exposure and thus its success. Formality can be defined as the disambiguation of communication; meaning a formal statement will *"… mean the same at different times, in different situations, or [when used by] by different people."* (Heylighen & Dewaele, 1999, p.24). The purpose of this study itself was three-fold. For one, naturally, we investigated the effect that formality has on science communication; however, the study also marks one of the first applications of transcript analysis within the empirical domain of YouTube content. We also aimed to expand on the understanding of formality as measure itself. What was found is that, of the five measures tested, the most robust measure of formality, the F-Score, does indeed positively influence the success of science communication. It is advised that science communicators do not unnecessarily deliver information in a less formal manner. However, this is nuanced by the finding that formality decreases the Engagement Rate ever so slightly, which means that the effect may only work positively on the short term; however, this phenomenon demands further studying. Aside from the actual effect, several limitations to the method were found. First and foremost, we found that computer generated transcripts are lacking in accuracy which hurts the results. Secondly, we advise future research to step beyond the most popular channels for a more robust analysis. Third, we emphasize that our claims on popularity do not equate to information retention or effectiveness. Fourth, future research must aim to retrieve a more holistic depiction of video presentation, a video is typically much more than just what is said within it. And finally, the current Natural Language Processing (NLP) tools demand further development in order for them to be truly viable for tasks as complex as identifying formality. Future researchers must expect to contribute new methods to the popular libraries.

Keywords: *Science Communication, Formality, YouTube, Transcript Analysis, NLP*

ACKNOWLEDGEMENTS

With this work I will be concluding seven consecutive years of intellectual and, perhaps most importantly, personal development. Throughout the years, I have been able to find what I love spending my time on and explore the various topics that keep me curious. I have been very fortunate to have been offered such an opportunity and I view it as an immense privilege.

However, I could not have done it without the support from those around me and those close to me. The first person that needs addressing would have to be my father. He has supported me throughout my educational journey by covering the costs of both tuition and books, enabling me to study and offering me the freedom to choose whichever major I so desired. The other two people I would like the thank are my brother and my best friend, who have stood by me both through personal as well as financial hardships; always ready to lend a hand. Last, but certainly not least, on a personal level, I would like to thank my dear girlfriend Zoi who has been motivating and supporting me throughout the entirety of my MSc. studies.

Academically there is also lots of gratitude to extend. First and foremost, I would like to thank my Thesis Supervisor Dr. Dainis K. Zegners who has been the person off which I could always bounce off ideas and who has also shown far reaching compassion for my personal circumstances that inhibited the timely delivery of various deliverables throughout the thesis process. In addition, I would like to thank Y. Lamrani Abou Elassad for his guidance during the completion of my premaster thesis. I have been able to apply many of the lessons he taught me on academic writing in the finalizing of this work. Next, I would like to extend my gratitude to a few of my classmates who have worked with me closely and with whom I was able to discuss various questions that arose throughout the thesis process. Finally, I must not forget the many lecturers at RSM who have inspired the route that I took with this thesis and helped me develop the academic foundation to successfully do so. Who also deserve mentioning are the kind people on StackOverflow.com that have helped me surmount multiple coding-related hurdles.

There are many more people that have helped me throughout the last challenging seven years that haven't been mentioned above; however, herewith I would like to thank anyone involved in making my education a success. The last seven years have seen some of the most interesting, difficult, exciting and incredible experiences I have had throughout my life, and I will happily and proudly look back on each and every one of them.

Maasdam, July 18th, 2021. *Gerbrand Laurens van Dijk*

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

1 INTRODUCTION

"It is the duty and the privilege of the well-informed citizen in a democratic society to make his private opinion prevail over the public opinion of the man on the street."
– Alfred Schutz's (1946) essay on the well-informed Citizen. –

A well-informed citizen is one that does not take everyday knowledge as a given; it means combatting ignorance despite not having expert knowledge on all the remotely relevant. As Schutz (1946) put it *"On the one hand, he neither is, nor aims at being, possessed of expert knowledge; on the other, he does not acquiesce in the fundamental vagueness of a mere recipe knowledge or in the irrationality of his unclarified passion."* (p. 466). Despite being able to function in ignorance, too, being well-informed enables us to be more independent, freer and discern between the true and false (Schutz, 1946). Augmenting our scientific literacy is one of the ways to bring us closer to that well-informed ideal (Sheufele and Krause, 2019).

1.1 Scientific Literacy, Communication and Formality

The diffusion of information has never been as *easy* as it is today. Nonetheless, in a world where information is at our fingertips, discerning truth from fiction has never been more *difficult*. Sheufele and Krause (2019) define misinformation as any information that is incorrect, either purposefully so or not. The proliferation of such misinformation has been linked to lacking scientific literacy as being one of the major contributors (Sheufele and Krause, 2019).

Scientific literacy as a concept, however, is often seen as diffuse and ill-defined (Champagne & Lovitts, 1989). The work of Laugksch (1999) explores the concept in great detail. The author ultimately settles on defining scientific literacy as a concept that is multi-dimensional and can both absolute as well as relative. Scientific literacy, as a definition, can range from the ability to relate scientific knowledge relevant to one's societal context to possessing over a certain set body of (advanced) scientific knowledge and skills. Burns et al. (2003) attempted to define scientific literacy within a single sentence: *"Scientific literacy is the ideal situation where people are aware of, interested and involved in, form opinions about, and seek to understand science."* (p. 190, cursive present in original source)

Science communication has the potential of playing a vital role in the advancement of public scientific literacy and the combatting the negative consequences of misinformation. In a

study by Ryder (2001) science communication was found to be a strong contributor in improving individuals' scientific literacy, this finding is also in line with what Burns (2003) and Kawamoto et al. (2013) found. The study that by Kawamoto et al. (2013) also found that, albeit unsurprisingly, scientific literacy is higher among people with previous scientific knowledge and/or experience.

The current state of science communication is often criticized for a variety of reasons, including its unengaging delivery and often poor presentation (Ashwell, 2016; Bonney et al. 2009; Riesch, 2015). YouTube, as a medium, lends itself particularly well to combatting current issues and furthering scientific literacy with science communication. For one, science education has long greatly benefitted from visual learning (Dimopoulos et al., 2003). Secondly, social media is known to be an effective tool for disseminating scientific information (Lamb et al. 2018). It is, therefore, also not entirely surprising that YouTube is already a staple in education (Jaffar, 2012). A study by Barry et al. (2016) found that, out of their sample of medical students, 78% used YouTube as their primary source of anatomy-related video clips. The popularity of YouTube as a medium for learning leaves it a great candidate for science communication.

The formalization of verbal expression concerns the disambiguation of the manner in which information is exchanged (Heylighen & Dewaele, 1999). The nature of the relationship between formality and science communication (if existent at all) is ill-researched. However, aside from sheer intuition, clues for there being a relationship do exist. On the one hand, ambiguity has been identified to negatively correlate with learning achievement (Carr et al., 2010), something which formality induced disambiguation would combat. On the other hand, formality has also been correlated with increased perceived social distance and reduced social credibility (Campbell & Wright, 2002; Irvine, 1979); which would be problematic since Welbourne & Grant (2016) identified the personal connection User Generated Content (UGC) provides to be particularly decisive to the popularity of science communication content on YouTube. Its unknown-yet-presumable effect and relatively high quantifiability makes formality a very suitable variable to assess within a novel empirical environment such as transcript analysis of science communication content on YouTube.

Considering the current state of the literature surrounding the topics introduced above and the potential contribution a study in such direction would have (detailed in 1.2), we arrive at the following research question:

RQ: *"How and to what degree does formality, as explored through transcript analysis, explain the performance of videos that aim to increase public scientific awareness?"*

1.2 Related Work

The present study combines natural language processing methods with the fields of online audio-visual content creation, science communication and sociolinguistic research into the aspect of formality. The academic relevance of this work is, therefore, three-fold—which we shall elaborate below.

Foremost, we will dive into the application of transcript analysis in an effort to explain YouTube video success. Whereas much has been researched on the influence of various meta data metrics on the success of YouTube videos, far less is known on what creators can do to improve their content itself. Variables such as video age, duration, engagement, viewer & uploader demographics, geographic locality, upload schedule, various typologies and more have been extensively studied (e.g., Borghol et al., 2012; Brodersen et al., 2012; Chatzopoulou et al., 2010; Figueiredo et al., 2011; Pinto et al., 2013; Welbourne & Grant, 2016; Yu et al., 2015). Aside from such meta statistics, content analysis does exist in the form of audio-visual analysis. Here either the audio or visual quality is investigated (e.g., Newman, 2018), or—as was done in the study by Cheng et al. (2013)—a great selection of audio-visual cues are combined in an effort to predict video success. Another category of studies is those where content features were encoded manually by the researching team themselves (e.g., Hussin et al., 2011; Keelan et al., 2007; Paek et al., 2010; Stellefson et al., 2014; Waters & Jones, 2011). A thorough review of the literature revealed that transcript analysis through the application of modern Natural Language Processing techniques, on the other hand, is yet to be explored and may thus yield novel fundamentalistic insights into the utility of the method itself.

Next in order, the academic community concerned with scientific literacy has long expressed the desire to improve on science communication (Burns et al., 2003; Cole & Cole 1968; Eisenhart, 1996), as it can be a strong tool in furthering scientific literacy (Ryder, 2011). A study by Welbourne & Grant (2016) investigated the influence of the video's presentation on the success of science communication on YouTube. This particular study shares the greatest resemblance with the present study, and our study adds to this work by exploring the effect that different measures of formality have on the success of science communication on YouTube through the application of transcript analysis. Formality fits in well as a presumable contributing variable due to the need for disambiguation in educational context (Carr et al., 2010) and the negative effect that informality is proven to have on the success of science communication (Riesch, 2015).

Finally, there is the matter of communicative formality itself. This study draws its definition for formality from two main sources: namely, Heylighen & Dewaele (1999) and Irvine (1979). The definition of formality is convoluted and used with differing intentions and in strongly varying contexts. When looking that the concept from a broader point of view it becomes clear formality can be measured through various other linguistic measures (further detailed in section 2.3.3). This study will explore the merit of each of those and the novel context may lead to renewed conceptual understanding of formality itself.

1.3 Scope & Next Chapters

This study investigates three focal topics: science communication, formality and transcript analysis. Science communication will be defined and scoped based on the definition put forward by Burns et al. (2003) in section 2.2.1. Formality will follow the definitions as elaborated by Irvine (1979) and Heylighen & Dewaele (1999), detailed in section 2.3.1. Only YouTube content within the science and technology genre will be considered, and each of the videos included will need to satisfy the selection criteria set in section 4.2. Hence, the study will not be generalizable outside of that genre. In addition, generalizability will halter at the most popular science communicators on YouTube since their channels ought to be considered outliers in terms of the whole field. In terms of transcript analysis, we will be considering the analysis of computer-generated transcripts. This is a decidedly different discipline from transcripts generated by humans.

The document is structured as followed, the subsequent chapter (2) will feature a detailed literature review into the topics of popularity on YouTube, science communication, formality in communication, and the differences between spoken and written communication. Succeeding will be a chapter (3) detailing the theory posed within the present study, its overarching research hypothesis and the conceptual framework. Next the method of data collection, processing and analysis will be explained in full within chapter 4, in order to ultimately arrive at the results in chapter 5. Concluding, chapter 6 will offer a discussion of all the findings and the method itself, and in chapter 7 we will go over some of the limitations the study is subject to and some recommendations for future research. The references are found at the end of the document in chapter 8 as well as the appendix in chapter 9.

2 LITERATURE REVIEW

Within the following literature review, an overview of the academic field will be offered for four distinct topics: namely, (2.1) popularity on YouTube, (2.2) science communication, (2.3) formality in communication and (2.4) the differences between spoken and written communication.

2.1 Popularity on YouTube

YouTube has facilitated the diffusion of user and/or professionally generated audio-visual content ever since February 2005 (YouTube, n.d.a). The content uploaded to the platform comes in many different genres; within their API, YouTube distinguish between 32 distinct video category ids (YouTube, n.d.b). The present study concerns video category id number 28: Science and Technology, of which science communication is a subgenre.

Of all popular Video on Demand (VoD) services available today, YouTube is by far the most social one and, unsurprisingly, appealing to the platform's social nature can greatly influence the success of a given video. A study by Susarla et al. (2012) took exactly this network perspective and concluded that the strategic connection of a channel with its fellow participants within the network has a positive influence on the popularity of its videos. The authors explain their findings through the enhanced ability of these channels to attract search traffic towards themselves. The sheer volume of content on the platform renders the power of search keywords much weaker for content discovery than that of recommendations within the users' subscriber networks (Susarla et al., 2012). This finding also further substantiates the "Rich-get-richer" effect of the YouTube content discovery system, as identified by Borghol et al. (2012). Borghol et al. (2012) also introduce several content-agnostic factors that may play into the popularity of YouTube content in general. The authors found the video's age, its average rating and the total number of likes awarded to the video to be best content-agnostic predictors of its success.

One of the ways in which we can distinguish YouTube content would be on the basis of whether it is created professionally (Professionally Generated Content; PGC) or by users of the platform itself (User Generated Content; UGC). Welbourne & Grant (2016) investigated the difference between either type of content in the context of science communication videos in particular and concluded that UGC content was significantly more popular than PGC content. The authors justify their findings by stating the importance of trust when consulting scientific resources; channels that produce user generated content often feature a greater

continuity in hosts which, in turn, facilitates the growth of trust between the viewer and the outlet (Welbourne & Grant, 2016).

This same study also found the rate of speech to be a significant predictor of the success of science communication content on YouTube (Welbourne & Grant, 2016). This is surprising, since typically any content that is intended to be educational, benefits from a slower pace of delivery to facilitate comprehension (Griffiths, 1992; Tauroza and Allison, 1990). The authors explain their finding as being characteristic for YouTube as a medium. The fact that videos can be played back means that anything that is not comprehended initially can easily be repeated and the additional density of easily interpretable sections leads to time economies for viewers (Welbourne & Grant, 2016).

Whereas the findings by Welbourne & Grant (2016) suggest that production value is not leading in determining the popularity of science communication content on YouTube, there are indications that there is a certain qualitative threshold that content ought to pass. In an experiment by Newman & Schwartz (2018) participants were presented with identical videos with varying audio quality, those exposed to poor audio quality assessed the content as significantly less favourable and less credible. In addition, Dobrian et al. (2011) found that buffering time was a strong predictor of video popularity (across various genres). All of which allude to a certain expected (material) quality of viewer experience.

When it comes to the literal measurement, success on YouTube can be measured in terms of both engagement as well as View Count (Susarla et al., 2012). Where the latter cannot logically be measured in any form other than the exact metric itself, the former is operationalised differently by the different authors that study the participation on YouTube's platform. Most choose to measure several engagement metrics independently (e.g., Chatzopoulou et al., 2010; Cheng et al., 2013; Figueiredo et al., 2011; Hussin et al., 2011; Paek et al., 2011) others choose to integrate these into one Engagement Rate (Goobie et al., 2019; Welbourne & Grant, 2016). The metrics tracked by those that measure them independently include Number of Shares, Number of Likes/Dislikes, Number of Comments and the Number of favourites. Several renditions of Engagement Rates exist as well. Welbourne & Grant (2016), for example, operationalize it as engagement events per view, each of the engagement metrics are divided by the total number of views. However, Goobie et al. (2019) chose to integrate all of these in one all-encompassing measure by summing the values of each of the metrics before dividing it by the total number of views. Whereas some differences exist in terms of how engagement is assessed, thorough consensus exists on View Count being the most accurate

measure of popularity; every single work considered for the literature review that investigated popularity on YouTube made use of this metric.

2.2 Science Communication

One of the key concepts studied within the present study is Science Communication. Within the following section we shall make an effort to define it and elaborate on its connection to the concept of scientific literacy and reason on its importance.

2.2.1 Defining the Concept

Within the realm of efforts that surround the communication of scientific findings and/or practices the variety of similar yet distinct definitions/concepts used concurrently make for a seemingly diffuse and ill-defined field of research (Champagne & Lovitts, 1989). Three terms used in conjunction particularly often are: science communication, scientific awareness and scientific literacy.

The first of which, Science Communication, is defined by Burns et al. (2003) as *"… the use of appropriate skills, media, activities, and dialogue to produce one or more of the following personal responses to science: Awareness, Enjoyment, Interest, Opinions or Understanding"* (p. 191). Science communication comes in different forms. Prior to the emergence of web 2.0, science communication was primarily done by scientists themselves (Valenti, 1999); however, nowadays the consumer of science information is increasingly turning to online resources (Brossard, 2013). Scientific awareness—on the other hand—as per the definition presented by Burns et al. (2003), is no more than a part of the premise of science communication. The authors define the premise of Science Communication to be four-dimensional: in the first place the aim is to (1) reach public awareness and (2) a general understanding of science, (3) to then work towards so-called "scientific literacy" (explicated in the next section) and ultimately (4) cultivating science culture where one can speak of a *"… society-wide environment that appreciates and supports science and scientific literacy."* (Burns et al., 2003, p. 190).

2.2.2 Scientific Literacy

Within the work by Burns et al. (2003), and consistent with the work of their fellows (Eisenhart et al., 1996; Norris & Philips, 2002; Matias et al., 2020), scientific literacy takes centre stage in answering the question of why we engage in science communication in the first place; hence, it is important to understand what scientific literacy is. Disagreement exists among

academics regarding an all-encompassing definition of scientific literacy. Some choose to define scientific literacy as a more general concept referring to what the general public should know about science in order to be productive citizens (Durant, 1993; Deboer, 2000; Eisenhart, 1996). Other's critique this viewpoint by stating it is "deceptively simple" (Laugksch, 1999, p. 90), or excessively generalized and disregarding the literal interpretation of literacy therewith disregarding important facets of the concept such as the ability to read and comprehend scientific text (Norris & Philips, 2002). Norris & Philips (2002) also put forward the most compelling and complete definition (and the one followed within this study) of Scientific literacy, adapted from the work of Rutherford & Ahlgren (1990) in their publication *Science for all Americans*:

"Scientific literacy—which encompasses mathematics and technology as well as the natural and social sciences—has many facets. These include being familiar with the natural world and respecting its unity; being aware of some of the important ways in which mathematics, technology, and the sciences depend upon one another; understanding some of the key concepts and principles of science; having a capacity for scientific ways of thinking; knowing that science, mathematics, and technology are human enterprises; and knowing what that implies about their strengths and limitations; and being able to use scientific knowledge and ways of thinking for personal and social purposes." (Rutherford & Ahlgren, 1990, as cited in Norris & Philips, 2002)

Improving scientific literacy through effective communication has many benefits such as improving inclusion, participation and academic performance. Miller (1998) emphasizes the importance of scientific literacy in the context of policy making and the increasing need for policies that feature a significant number of technical and scientific details. However, scientific literacy does not only cover the ability of applying detailed scientific knowledge to practical situations, but there is also a civic scientific literacy that airs more on the side of awareness (Kawamoto et al., 2013). YouTube, as a medium, lends itself particularly well to furthering this type of scientific literacy due to the accessibility of its content, which is crucial since it is often those hardest to reach that can use an improvement in their scientific literacy the most (Matias et al., 2020).

2.3 Formality in Communication

A second key concept within the present study is that of formality, namely formality in communication. We shall elaborate on how the concept is defined in the context of this study, the determinants of its prominence and how we could attempt to measure it.

2.3.1 Defining the Concept

The concept of formality is rather convoluted due to it being abstract yet also intuitively understood by many. Within a study by Pavlick and Tetreault (2016), a substantial interclass agreement (± 0.6) was found when asking humans (via Mechanical Turk) to rate 1.000 distinct sentences in terms of formality. However, whereas humans do display ample coherent understanding of what formality comprises of, its convoluted nature remains to hamper quantitative measurement (Pavlick & Tetreault, 2016).

A well-respected work within the field of socio linguistic research on the topic of formality is that of Judith Irvine (1979). She lists 4 distinct components that distinguish the formal from informal. The first of which is increased code structuring, referring to an overall increase in discourse rigidity; both linguistically and with regards to other elements of social code (Irvine, 1979). The author proceeds to list the second item as code consistency: the act of having different social codes signifying the same thing (e.g., tone and choice of words). The third element Irvine (1979) identifies is the invocation of positional identities. She explains that this refers to the level of acknowledgement among conversing parties that differences in social rank exist; a more formal situation usually features stronger acknowledgement of said differences. The emergence of a central situational focus is listed by Irvine (1979) as a fourth and final element alluding to a more formal situational context. In a more formal context, the topics up for discussion are distributed with less leniency; one is expected to more or less stick to the main implied talking point (Irvine, 1979). Irvine's (1979) work homes in on the situational aspects of formality but leaves the concept unmeasurable.

Heylighen (1999) makes an attempt at resolving this and arguably primarily builds upon the first element listed by Irvine (1979). Within their work the author deems an expression to be "formal" when it has "invariant meaning" (Heylighen, 1999). The variance Heylighen (1999) refers to is strongly connected to the concept of deixis, meaning the degree of context dependency of a given expression. This is still vague to some degree due to the obscureness of terms such as meaning and context; however, Heylichen (1999) further elaborates that formal expressions ought to *"… mean the same at different times, in different situations, or [when used by] by different people."* (p. 27) and that meaning refers to the ability of an expression to recognise some abstract entities as being one thing and distinguish them from other entities that they are not. To further explicate the latter, Heylichen, together with his colleague Jean Marc Deweale, propose the following example: *"A prototype of formal language might be the sentence read out by a judge*

at the end of a trial. Prototypical informal speech would be produced in a relaxed conversation among close friends or family members." (Heylichen & Deweale, 1999, p. 2)

2.3.2 Determinants

Having established that spoken formality is largely dependent on how much context is acceptable, one may already infer that situational context is highly important. Within a study by Bello (2005), for example, participants were situated in different contexts varying in terms of formality which led to them displaying various communicative strategies where people chose to equivocate more (i.e., communicate more ambiguously) as the setting becomes more formal. This contradicts the findings and explanations by Heylighen & Dewaele (1999) regarding the situational determination of communicative formality. The authors found that as individual interest intensify, so does a need to converse with clarity (Heylighen & Dewaele, 1999). The authors Heylighen & Dewaele (1999) go on to state that this—in turn—could explain why job interviews, for instance, tend to be more formal; the situation requires both parties to communicate as unambiguously as possible. The contradiction found within the literature leaves room for further investigation of the role of situational context in determining formality.

Aside from the context the speaker/writer finds themselves in, their personality also plays into the degree of formality their choice of words is likely characterized by. A large-scale literature review on language and personality by Adrian Furnham (1990) identified introversion, in particular, to be a strong predictor of communicative formality. The more introverted an individual is, the more likely they are to communicate with greater degrees of formality. A later study by Heylighen & Dewaele (2002) reinforces the latter with similar findings. A study by Mehl et al. (2012) also found similar results; however, the study identified an impactful moderating role for context here as well. The authors found that where extroverted people communicate very directly in public environments, they do not necessarily do so in more private environments. The inverse appears to be true for those displaying strong indications of the neuroticism personality trait (Mehl et al., 2012).

Level of education is another factor contributing to the degree of formality in communication. Francis & Heylighen (1999) identified the level of education of a given individual to influence the formality of how they choose to verbalize their ideas. The authors nuance their findings by saying this is most clearly the case for written communication and the evidence is weaker for spoken communication. This correlation could be explained by the, on average, richer vocabulary of those who enjoyed higher formal education which enables them to voice their thoughts more accurately and provides them the additional mental capacity

needed to communicate unambiguously (Heylighen & Dewaele, 1999). Interesting to note is that the same study found the parents' educational level to correlate negatively with the users' choice of formal verbalization. The authors justify their findings through the rather academic context of where the data was collected; those participants whose parents were highly educated likely felt more familiar with such an environment and thus felt less of a need to communicate utmost unambiguously.

2.3.3 Measurement

Formality, as defined above, can be measured (directly and indirectly) in several ways, including: the F-score (Heylichen & Dewaele, 1999), Readability scoring (Edgar & Chall, 1949), Lexical Density (Halliday, 1985) and Syntactic Complexity (Nippold, 2017). The following section will detail a description on each of these metrics.

Following from the definition put forward in section 2.3.1, Heylichen & Dewaele (1999) actually define two different kinds of formality: surface-and-deep formality. Deep formality largely but also exclusively conforms to the definition posed above; namely, only non-deictic expressions can be considered truly to be formal. However, lexical deixis is difficult and time consuming to measure and the authors pose that there is also a surface level formality to be discovered that can actually be measured. For this purpose, the authors devised their F-score which takes into account the frequency of formal parts of speech such as nouns, adjectives, articles and prepositions, against the frequency of informal ones such as pronouns, verbs, adverbs and interjections.

$$F = \left(\frac{\begin{array}{c} Noun\ freq. + Adjective\ freq. + \ + \text{ preposition freq.} + \text{article freq.} \\ - \text{ pronoun freq.} - \text{ verb freq.} - \text{ adverb freq.} - \text{ interjection freq.} + 100 \end{array}}{2} \right)$$

Equation 2.1 F-Score Formula

Readability studies the ease with which a text can be comprehended, this goes beyond the aspect of legibility and is dependent on style, content and typography (Dale & Chall, 1949). The formula put forward by Dale & Chall (1949) reads as follows:

$$Readability = 0.1579 \left(\frac{Difficult\ words}{Words} \times 100\% \right) + 0.0496 \left(\frac{Words}{Sentences} \right)$$

Equation 2.2 Dale-Chall Readability Formula

It uses a list of 783 words that within their study were labelled as "simple", and all other words are considered "difficult". Readability scoring is particularly well suited to written documents and less suited to spoken passages.

Halliday's (1985) formula of determining lexical density may also inadvertently measure formality. It measures the number of lexical items as a percentage of the total number of clauses within a text.

$$Lexical\ Density = \left(\frac{Lexical\ items}{Total\ number\ of\ clauses} \times 100\%\right)$$

Equation 2.3 Lexical Density formula

Lexical items, in this case, are defined as either nouns, adjectives, verbs or adverbs (Halliday, 1985). This makes the formula fundamentally different than the two above in the sense that verbs and adverbs are included. The formula was meant as a discernment between verbal and written text; however, since verbal text is often less formal it may also act as a measure of formality (Heylichen & Dewaele, 1999).

Related to the aforementioned is syntactic complexity; a measure primarily reflecting the density of information and the complexity of one's choice of words. The measure was pioneered by Nippold and her colleagues (2017) and measured using a combination of two measures, namely: mean length of communication units (MLCU) and clausal density (CD).

$$MLCU = \frac{1}{n}\sum_{i=1}^{n} LCU_i = \frac{LCU_1 + LCU_2 + \cdots + LCU_n}{n}$$

Equation 2.4 Mean Length Communication Unit formula

&

$$CD = \frac{1}{n_{cu}}\sum_{i=1}^{n_{cu}} Clause_i = \frac{CLause_1 + Clause_2 + \cdots + Clause_n}{n_{cu}}$$

Equation 2.5 Clausal Density Formula

A communication unit is *"… an utterance that contains a main clause; it also may contain one or more subordinate clauses that are attached to it; usually a sentence"* (Nippold et al., 2014, p. 1346). The MLCU therewith measures the lengthiness of expression, which we know is positively correlated with formality (Irvine, 1979). Clausal density, on the other hand, measures how dense the information is, and we also know that written communication is typically highly dense (Chafe, 1982) but also much more formal (Heylighen & Dewaele, 1999).

2.4 Spoken and Written Communication

Most text analysis methods outlined within this chapter have found their origin through the analysis of written communication. Within this study we analyse transcripts of spoken passages and one cannot simply assume written and spoken language can be treated as equal. Horowitz & Newman (1964) investigated the difference between both modes of discourse and concluded that there exists a plethora of dissimilarities between spoken and written communication. First of all, the authors describe spoken communication as "profligate"– meaning wasteful; while conversing one tends to use more words yet convey fewer ideas as compared with written forms of communication. Spoken communication also tends to be far more repetitive than written communication and produce much more subordinate, or "elaborative", material (Horowitz & Newman, 1964).

Chafe (1982) elaborated on the differences above by agreeing with the former in that speaking is indeed "easier" than writing, which the author explains by stating how when writing our thoughts must always be ahead of their expression; leading one to think much more of what one expresses. Written language can be said to be "integrated", in the sense that sentences tend to comprise of multiply ideas as one whole. Whereas spoken language—on the other hand—can be said to be fragmented in sentences that do not convey much more than a single idea (Chafe, 1982). Moreover, Chafe (1982) adds one crucial aspect to the theory; namely, speakers interact with their audiences whereas writers do not. This brings about two key elements in which either mode differs from one another which the author calls detachment and involvement. Written language is characterized by detachment, writers tend to use the passive voice much more often than speakers since they and their audience are detached in space and time. Spoken language, on the other hand, is characterized by involvement; speakers share much more information with their audience and in doing so will monitor the flow of information much more closely. Speakers will speak in the first person more often and give many personal (at times "fuzzy") examples (Chafe, 1982).

The presumed difference between the two modes is not uncontested, however. In a study by Akinnaso (1985), the author criticizes the literature by stating the difference in language use is much less apparent when not exclusively looking at academic writing and identifies formality as a confounding variable. Halliday (1985), too, states the inherent similarities between the two modes by saying spoken English is no less structured than written English principally speaking since *"… both [are still] 'kinds of English', and the greater part of their patterning is exactly the same."* (p. 79).

3 THEORY

3.1 Theoretical Framework

Within in this chapter we devise our research hypothesis following the research question as posed in section 1.1. On the whole, we hypothesize there being a relationship between formality and the success of science communication; however, as elaborated in 1.1, due to conflicting findings within the academic field we do not presume anything regarding the strength or nature of said relationship. Hence, the overarching research hypothesis (ORH) studied within the present study reads as followed:

ORH: *A relationship exists between the degree of formality within a science communication video's transcript and the success of that video on YouTube.*

This will be the main and only hypothesis we will be studying throughout the research. However, we will be approaching it from different points of view. As we have seen throughout the literature review, communicative formality has many different facets to it. Its strong relationship with situational context makes it very complex. Said situations can be characterised in many different ways as shown throughout the work of Irvine (1979), which—in turn—means there are also many indicators for formality. This leaves us with various ways to measure formality, as was also explored in section 2.3. Throughout our study we will be exploring the different measures of formality and their merit specifically in the context of science communication. Each measure can be said to home in on a different facet of the concept of formality and comparing the contribution of each of the measures has left us enabled to characterize the influence of the variable more deeply.

The first measure elaborated within the literature study is that of the F-score as devised per Heylighen & Dewaele (1999). Within their work, the authors point to ambiguity (or rather the lack thereof) to be the strongest indicator of communicative formality. They distinguish between two kinds of formality: deep formality and surface level formality (Heylighen & Dewaele, 1999). The authors have devised a measure for formality, namely F-score, which is meant to measure surface formality within any given passage. The F-Score assesses the form of a passage by enumerating the presence of parts-of-speech that allude to ambiguity. In the study by Welbourne & Grant (2016) pace of speech was identified as being one of the major variables

predicting the success of science communication content; hence it holds particular promise within the present study.

When making inferences on the nature of the relationships we expect to observe it is important to take into account the format of our data. As posed within the literature study, spoken and written communication cannot be treated as equal outright. The data used within this study will reflect transcripts of texts that were meant to be spoken but have likely been scripted before; determining how the data ought to be classified then becomes difficult. However, according to the work by Welbourne & Grant (2016) engaging the audience is important in creating successful YouTube videos, especially those within the science communication genre, and the work by Chafe (1982) implies that spoken language is far better at establishing that kind of connection. Assuming that over time creators learn what leads to more successful content (Arthurs et al., 2018), it is conceivable that the majority of videos in our will have been prepared as spoken language.

Readability, as a second way of measuring formality, is a measure typically meant for written texts (Dale & Chall, 1949). Aside from its premise in measure formality, the inclusion of readability as a variable will serve a second purpose as well, namely, to characterize our data. The strength and direction of the relationship we observe can inform us in how we may approach transcript analysis on YouTube. In case the variable shows a strong and statistically significant contribution we may at least conclude that transcripts and the science communication content itself can be approached as written texts. The direction, in turn, will inform us on how formal the text is likely to be, with more readable texts typically being associated with lower degrees of formality.

The next measure proposed within the literature study is that of lexical density as defined by Halliday (1985). Lexical density is also a measure well suited to discerning spoken language from written language; however, it takes a different perspective from readability. Where readability assumes a text is intended as written communication, lexical density does not make such an assumption. Readability measures the *"difficulty"* of what is written, where lexical density measures how densely the information is packed within the spoken/written passage. Here spoken text is typically less dense than written text; however, in the context of this study we do not expect lexical density to have a very strong contribution as such since the similar format won't allow for as much variation within the variable. What it will encode however, is the pace of each video and how much information is relayed within each clause, which inadvertently will measure formality since formal sentences often contain more information (Heylichen & Dewaele, 1999).

Formal language is often more compact yet uses longer words (Heylighen & Dewaele, 1999). Mean Length Communication Unit (MLCU) is a robust measurement of lengthiness of expression, by both taking into account the amount of information being conveyed but also the manner in which it is being conveyed (Nippold et al., 2017). This measure approaches formality from a similar viewpoint as the F-score; namely, assessing its form of expression. However, contrary to the F-score, we are less interested in implicitly measuring the code of expression through analysing the parts of speech; rather we assess the explicit code of discourse used. Similar to lexical density, we would anticipate lengthier words to encode more information, more efficient, and thus disambiguate the passage they are contained within. This would consequently function as a measure of formality; lengthier expression equals more formal expression.

The last variable resulting from our literature study is clausal density (CD). CD is different from lexical density in that it does not reflect with how many words clauses are being presented, but rather it measures how many clauses are included within each of the communication units. The authors had operationalised that variable as words per minute. Clausal Density will work similarly within the present study except for one slight adjustment instead of words per minute it will take into account the actual ideas expressed through those words, as clauses per communication units.

3.2 Conceptual Model

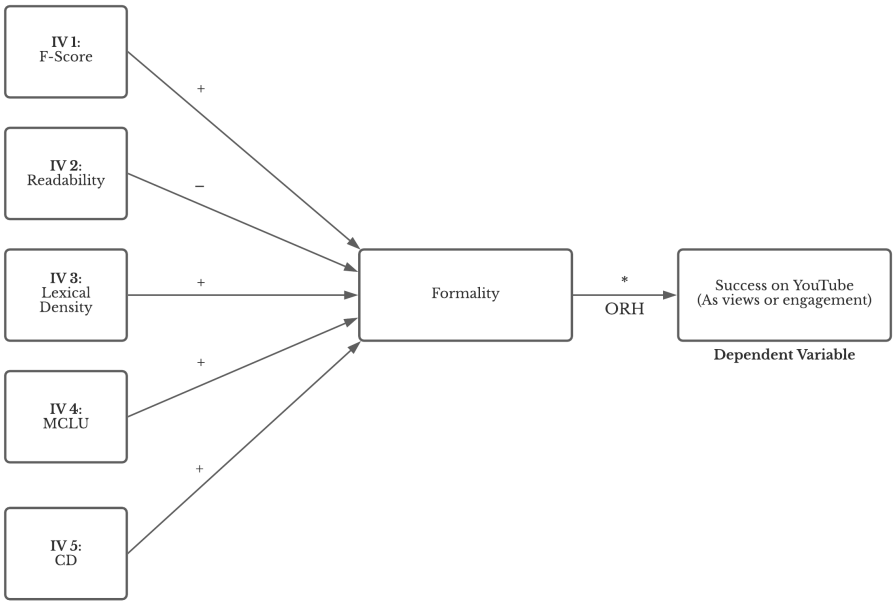


Figure 3.1 Conceptual Model

4 METHOD

The following section introduces and elaborates on the methods used to accrue and analyse the data used within our study. The chapter opens with a section on the overall research design (4.1) to then continue with a section on the data collection procedures (4.2). Since the data collection procedures are one of the major contributions of the study, we describe all 3 steps in great detail (4.2.1 - 4.2.3). Consecutively, in section 4.3, we provide the operationalization of our independent (4.3.1) and dependent (4.3.2) constructs as well as the various control variables (4.3.3) we intend to include. Following, we define the various empirical models (4.4) that are used to estimate the effects theorized above. Ultimately, we conclude with offering an overview of the resulting data matrix (4.5).

4.1 Research Design

Exploring the relationship we hypothesize, required the collection of a set of videos with differing degrees of success and differing degrees of formality. Consequently, the unit of analysis would be defined as the YouTube videos themselves (more specifically; their transcripts). This led to a study typically characterised as being cross-sectional, since we set out to elaborate on between-group differences.

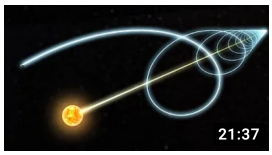
With the various empirical approaches further outlined in section 4.4 we hope to gain an understanding of two aspects. For one, we aim to understand within-channel differences to build upon the study of Welbourne & Grant (2016). However, since such a study inherently introduces researcher bias, we also devised a fixed effects model where the channel differences are negated to gain a better idea of the effects we can expect within a broader population.

4.2 Data Collection Procedures (3-Steps)

4.2.1 Step 1: Sampling

The selection of videos was done in adherence to the definition of Science Communication put forward by Burns et al. (2003): *"… the use of appropriate skills, media, activities, and dialogue to produce one or more of the following personal responses to science: Awareness, Enjoyment, Interest, Opinions or Understanding"* (p. 191). Five channels that had at least five videos posted at time of sampling which adhere to this definition of scientific literacy have been selected accordingly, a description of these channels is provided in table 4.1.

Table 4.1 Sampled Channels

Channel Name	Wikipedia Description	Sample Thumbnail
Veritasium	"In January 2011, Derek Muller created the educational science channel Veritasium on YouTube, the focus of which is "addressing counter-intuitive concepts in science, usually beginning by discussing ideas with members of the public"." [1]	
Vsauce	"Vsauce is a YouTube brand created by Internet celebrity Michael Stevens. The channels feature videos on scientific, psychological, mathematical, and philosophical topics, as well as gaming, technology, popular culture, and other general interest subjects." [2]	
Kurzgesagt	"Kurzgesagt (German for "In a nutshell") is a German animation studio founded by Philipp Dettmer. The studio's YouTube channel focuses on minimalist animated educational content, using the flat design style. It discusses scientific, technological, political, philosophical and psychological subjects." [3]	
Mark Rober	"Mark B. Rober (born 1980 or 1981) is an American YouTuber, engineer and inventor. He is known for his YouTube videos on popular science and do-it-yourself gadgets. Before YouTube, Rober was an engineer with NASA for 9 years where he spent seven of those years working on the Curiosity rover at NASA's Jet Propulsion Laboratory." [4]	
AsapSCIENCE	"AsapScience, stylized as AsapSCIENCE, is a YouTube channel created by Canadian YouTubers Mitchell "Mitch" Moffit and Gregory Brown. The channel produces weekly videos that touch on many different topics of science." [5]	

Source:

[1] https://en.wikipedia.org/wiki/Derek_Muller

[2] https://en.wikipedia.org/wiki/Vsauce

[3] https://en.wikipedia.org/wiki/Kurzgesagt

[4] https://en.wikipedia.org/wiki/Mark_Rober

[5] https://en.wikipedia.org/wiki/AsapScience

From the channels listed above, we retrieved all videos that concur with the Science Communication definition. The YouTube API allowed us to do so via its `.search()` method, details on its configuration and a code example are provided in appendix 9.1.1. The selection's adherence to the definition was verified through manually judging 30 videos on their content at each sampling iteration. Each encounter of a wrongful inclusion would trigger a new iteration. Keywords that were found within video titles indicative of wrongful inclusion were listed and used as filtering criteria until the sample of 30 videos was clear of wrongful selections. In total, 8 iterations were required to exclude a total of 27 videos across all five channels.

Following the sampling of the complete channels was the creation of a random sample of 200 videos (when available). Random sampling allows us to mitigate some of the selection bias after having excluded certain videos. From each video three features were captured: the video id (to be able to uniquely identify each video), its title and the name of the channel that published it. What resulted is a sample with the frequency distribution by channel as shown in table 4.2.

Table 4.2 Sample Frequency Distribution

	Initial Selection			*Raw Sample**	
Channel Name	Frequency	Percentage		Frequency	Percentage
VSauce	366	28.8%		200	24.1%
AsapSCIENCE	359	28.3%		200	24.1%
Veritasium	306	24.1%		200	24.1%
Kurzgesagt**	141	11.1%		141	17.1%
Mark Rober**	98	7.7%	+	92	11.1%
Total	*1270*	*100%*		*831*	*100%*

** Not representative of the final sample, more cleaning was necessary.*

*** Channel has not posted more than 200 videos, and thus its entire science communication contribution was included.*

4.2.2 Step 2: Statistics Retrieval

The next step in the process is the retrieval of video statistics. The initial sample as detailed above only included three features: the video id, video title and channel name. The `.search()` method is incapable of retrieving more relevant metrics; however, by using the `.videos()` method we can retrieve much more. An overview of the various metrics is provided in table 4.3 and the configuration of the API as well as a code example is provided in appendix 9.1.2. While querying the API, not every video returned the metrics we aimed for. Ultimately, a total of 12 videos returned missing values, these records were excluded from the final sample.

Table 4.3 Overview of Meta Statistics

Metric	Example	Explanation
Published At	2017-08-22 T00:38:04Z	The date at which a video was first published, reported in ISO format, later changed to regular datetime format.
Duration	PT5M21S	ISO time format, later changed to number of seconds.
Definition	hd/sd	Video quality indication in High/Standard definition.
View Count	34159442	Number of views a video has attained at time of sampling.
Like Count	414887	Number of likes awarded to a video at time of sampling.
Dislike Count	13031	Number of dislikes awarded to a video at time of sampling.
Favorite Count*	0	Number of favorites awarded to a video at time of sampling.
Comment Count	39069	Number of comments a video has received at time of sampling.

** This feature was later dropped since all videos returned 0.*

4.2.3 Step 3: Downloading & Transcribing

After having retrieved the meta data on the videos in the previous step, the next step was to access transcriptions of the videos. The transcription is necessary to make the videos "analysable". Unfortunately, the YouTube API has been subject to significant reductions in its capabilities due to privacy concerns and one of the changes that was made is that automatically generated captions cannot be accesses by those without the specific channel credentials. This meant the use of 3rd party services for data collection was unavoidable.

To download the audio tracks of the selected videos, the youtube-dl Python library was used. This library operates at the command line which was controlled through a subprocess in Python. The configuration of the downloader along with a code example can be found in appendix 9.1.3. Transcribing the audio tracks was done through means of the IBM Watson Speech-to-Text Service. This API allowed us to transcribe a large array of audio data at relatively low cost with acceptable accuracy (around 80% of a given track is entirely correctly transcribed). The configuration of the Speech-to-Text service is also provided in appendix 9.1.4 with accompanying code example. Both the youtube-dl library as well as the Speech-to-Text service from IBM-Watson had cases where no data could be retrieved, these instances were omitted from the final sample. The resulting final sample frequency distribution by channel is provided in table 4.4. On average 16.4% of data is lost through this approach.

Table 4.4 Final Sample Frequency Distribution

	Final Sample			*Data Loss from Raw Sample*	
Channel Name	Frequency	Percentage		Frequency	Δ Percentage
AsapSCIENCE	156	22.4%		-44	-22%
Veritasium	143	21.8%		-57	-28.5%
Kurzgesagt	139	21.2%		-2	-1%
VSauce	139	21.2%		-61	-30.5%
Mark Rober	87	13.3%	+	0	0%
Total	*664*	*100%*		*-164*	*16.4%**

** Average.*

4.3 Operationalisation

Within our study the dependent construct is the success of science communication content on YouTube and the independent construct is the formality of its delivery. The following sections goes into depth on how either construct is made measurable (as per chapter 3).

4.3.1 Independent Variables

4.3.1.1 F-Score

The formula put forward by Heylighen & Dewaele (1999) measures the prevalence of various Parts-of-Speech (POS). This meant the transcripts needed to be tokenized as well as tagged for POS. In Python, there are two main libraries that are used for Natural Language Processing tasks as such: NLTK and spaCy. NLTK was established in 2001 and has been expanded ever since; it is one of the most advanced and powerful libraries in existence for NLP tasks (Bird, 2009). spaCy, on the other hand, was built on top of NLTK and takes the most popular methods for each NLP task and bundles those up into one, highly efficient, library. Since this study does not demand the variety that NLTK offers, it was opted to use spaCy. The F-Score was calculated by using the spaCy pipeline to tokenize the transcripts and then extracting and counting the POS tags from the resulting spaCy documents.

4.3.1.2 Readability Scoring

There are many renditions of readability scores; however, as mentioned in chapter 3, this study will be following one of the oldest formulae, that by Dale and Chall (1949). The formula relies on a list of (now) 3000 words that are familiar to at least 80 percent of fourth-

graders which was extracted from readabilityformulas.com (Scott, 2021). Ultimately, the readability formula was defined by simply comparing every token in the document with the word list.

After tokenizing the documents, a next step often entails either removing stop words and stemming/lemmatizing the document; meaning reducing the words to their stem or lemma (e.g., stemming: running, ran, run → run, ran, run; lemmatization: going, went, gone → go). In case of this study, the choice was made not to stem, lemmatize nor remove stop words from the documents. This is because none of the features we define would benefit from it; actually, the contrary. A feature such as the readability score is meant to measure a natural representation of language, anything that deteriorates that naturality reduces the fit of the chosen measures to the purpose.

4.3.1.3 Lexical Density

For this measure, and the following, it was necessary to identify clauses within the text. Both independent (convey meaning on their own) as well as dependent clauses (refer to another independent clause) were identified. Doing so meant introducing a new NLP library; namely, Stanza. The Stanza library is an NLP library created by Stanford University which, crucially, contains the Stanford Dependency parser. This parser is capable of creating a tree model of a sentence to illustrate which words reside within/depend on which clauses. By counting the unique 'head' tags per sentence we were able to count the number of dependencies (dependent clauses) and identify independent clauses by counting 'root' tags. Lexical Density was then calculated using this newly enumerated clauses feature and the spaCy POS tagging feature.

4.3.1.4 MLCU & Clausal Density

The Mean Length Communication Unit and Clausal Density are closely related in that they both rely on the parsing of communication units. The Nippold (2017) paper was used to accurately define and parse communication units. Using the tagged communication units and clauses either of these features were defined by simply reconstructing their formulae in Python.

4.3.2 Dependent Variables

The dependent concept within the present study is typically measured in two possible ways: through measuring View Count or through measuring engagement. The former is the easiest; it is simply the number of views a video has accrued at a given point in time, such as we have extracted from the YouTube API in section 4.2.2. The latter, however, is typically defined

through a combination of metrics: namely, the number of ratings (likes and/or dislikes), comments and shares/favourites. However, as pointed out by Pinto (2013), it is not helpful to include the absolute measures within a linear regression model with View Count as a dependent variable since they are highly correlated. Hence, within the present study we will operationalize the Engagement Rate along the same lines as Goobie et al. (2019): we will take the sum of the ratings (likes and/or dislikes) and comments and divide them by the corresponding View Count in order to assess how many users that view the content also interact with it and we will create a separate estimator for this. The YouTube API does not allow us to retrieve the number of shares and for the number of favourites the API returned 0 for every video for unknown reasons.

4.3.3 Control Variables

Since success on YouTube has been abundantly, and fruitfully, predicted using different meta-statistics (e.g., Figueiredo, 2011; Goobie et al. 2019; Susarla et al., 2012; Welbourne & Grant, 2016), we will be collecting all available meta statistics the YouTube API allows us to query. In addition, it is sensible to control for the length of the various documents by including some text length statistics since they are likely to vary strongly.

However, not every metric poses as a good control variable. First of all, we exclude channel statistics, such as subscriber count and channel age, since they are controlled for by our Fixed Effects Model (detailed in the next section). Next, we also exclude word, character and sentence count since they cause multicollinearity as well. Average word length and average sentence length are more robust measures. An overview of the control variables is offered in table 4.5.

4.4 Empirical Models

To estimate the hypothesized relationships, we will deploy several Ordinary Least Squares (OLS) regression models. The various models are defined to estimate the two different success metrics (Models A: View Count & Models B: Engagement Rate). In addition, to gain insight into the individual contributions of each variable, the independent variables will be introduced stepwise. We devised two variations of models: regular OLS models and Fixed Effects models. The OLS Models show the effects within this particular group of popular

Table 4.5 Control Variables

Available Metrics	Control variables
Channel age	
Number of subscribers to channel	
Average views per channel	
Engagement Metrics	Video Age (in days old)
Video age	Video Duration (in seconds)
Video duration	Average Word Length (in characters)
Character count	Average Sentence Length (in words)
Word count	
Sentence count	
Average word length	
Average Sentence length	

science communication channels on YouTube. Whereas the Fixed Effects models show the effects when the channel effects are negated. We do so by subtracting the channel mean for each variable from their value. This created a demeaned dataset which essentially allows us to control for facts that are constant yet unobserved between channels (such as the presenter, the channel name, the channel age etc.). In addition, for the models based on View Counts, the dependent variable is transformed logarithmically, this is further explained in section 5.2.1. The final models[1] we have estimated are the following:

$$M_{OLSiA}: \log(Y_{VC}) = \alpha + \beta_1 \text{F-Score} + \beta_2 \text{Readability} + \beta_3 \text{LexicalDensity} + \beta_4 \text{MLCU} + \beta_5 \text{CD} + \beta_i \text{ControlVariable}_i + \varepsilon_i$$

$$M_{OLSiB}: Y_{ER} = \alpha + \beta_1 \text{F-Score} + \beta_2 \text{Readability} + \beta_3 \text{LexicalDensity} + \beta_4 \text{MLCU} + \beta_5 \text{CD} + \beta_i \text{ControlVariable}_i + \varepsilon_i$$

$$M_{FEMiA}: \log(Y_{VC}) = \alpha + \gamma_2 \sum_{i=1}^{N-1} \text{Channel}_i + \beta_1 \text{F-Score} + \beta_2 \text{Readability} + \beta_3 \text{LexicalDensity} + \beta_4 \text{MLCU} + \beta_5 \text{CD} + \beta_i \text{ControlVariable}_i + \varepsilon_i$$

$$M_{FEMiB}: Y_{ER} = \alpha + \gamma_2 \sum_{i=1}^{N-1} \text{Channel}_i + \beta_1 \text{F-Score} + \beta_2 \text{Readability} + \beta_3 \text{LexicalDensity} + \beta_4 \text{MLCU} + \beta_5 \text{CD} ++ \beta_i \text{ControlVariable}_i + \varepsilon_i$$

[1] Due to the stepwise introduction of our independent variables a 5-fold of the models above will exist upon implementation.

4.5 Data Matrix

Table 4.6 Data Matrix

Instances			*Construct X*			
Video.ID	Title	Channel Name	F Score	Readability Score	Lexical Density	MLCU
G10m2ZZRH4U	Total Solar Eclipse (2017)	Veritasium	499.5	9.85	24.2	52.98

Table 4.6 Data Matrix (continued)

Construct X cont.	*Control Variables*				*Construct Y*	
Clausal Density	Days old	Duration	Avg. Word Length	Avg. Sentence Length	View Count in mln.	Engagement Rate
34.8	35	2040	4.3	22.8	7.418197	5,73

5 RESULTS

The following sections consists of two main parts, sections 5.1 and 5.2. Within the former section, we shall explore the data we have collected by first reporting on several descriptive statistics (5.1.1), after which we will explore the correlations between each of the analyzed variables (5.1.2). The section will be concluded with a look at the distribution of the dependent variables and their dependencies (5.1.3). Section 5.2 proceeds to discuss the regression results in vast detail. The section will start off with a piece on the necessary variable transformation steps (5.2.1) before the results per model variant are discussed (5.2.2–5.2.3).

5.1 Exploratory Data Analysis

5.1.1 Descriptive Statistics

Table 5.1 offers an overview of the descriptive statistics for the overall sample. As discussed in section 4.2.3, after cleaning the data we were left with 664 completed records. Several interesting findings can be distilled from the table below, such as that the most viewed video within the sample has over 115 million views and that all videos within our sample have been published between the 19th of June 2010 and the 21st of June 2021. We also find that the average video within our sample lasts 4 minutes and 57 seconds. The descriptive statistics are important to keep in mind when interpreting the regression results.

Table 5.1 Overall Sample Descriptives Statistics

	Count	Mean	Std.	Min	Median	Max
1. F-Score	664	124.53	69.59	-73	112	499.50
2. Readability Score	664	9.44	0.90	7,485	9.38	13.68
3. Lexical Density	664	41.52	43.25	1.9	23.40	199.50
4. MLCU	664	89.20	87.57	6,267	51.74	399.50
5. Clausal Density	664	24.92	25.13	1.34	18.9	198.67
6. Days old	664	2080.57	1174.71	1	2081	4,020
7. Duration	664	417.84	305.20	38	345	2,127
8. Like Count	664	198,382.49	300,247.65	247	87,620	2,339,474
9. Dislike Count	664	4,538.81	7,890.50	4	2,067	73,883
10. Comment Count	664	15,343.15	21,194.81	0	7,386	16,9280
11. Avg. Word Length	664	4.14	0.35	2.2	4.10	5
12. Avg. Sent. Length	664	20.28	8.71	5.7	18.60	94
13. View Count	664	8,481,121.47	12,856,530.2	28,311	4,814,907	115,540,672
14. Engagement Rate	664	2.71	1.72	0.371	2.31	10.79

5.1.2 Correlation Matrix

Table 5.2 shows the correlations between each of variables within our sample, and figure 5.1 offers a brief visualized overview of said correlations in the form of a correlogram. Overall, we find that there exist rather many significant correlations within our sample. We Will briefly touch upon the most intense and interesting findings for each variable, starting with the dependent variables before moving on to the independent variables and concluding with the control variables.

When looking at the variables that correlate with our first dependent variable, View Count, we find that they mostly consist of engagement metrics. As explained by Pinto (2013) this was to be expected, to post a like, dislike or comment a person first needs to view the video. This creates an inherent spurious correlation between the variables and forms the exact reason why we do not include the engagement metrics in a model that estimates View Count. For our second dependent variable, Engagement Rate, one particular finding jumps out: its correlation with duration (Engagement Rate ~ Duration: -0.709***). This finding potentially alludes to changes in the YouTube algorithm over time that creators have responded to.

Next, we consider the independent variables. First is the F-Score, for this measure we find it mostly correlates with the age of a video (F-Score ~ Days Old: -0.429***), the duration of a video (F-Score ~ Duration: 0.449***) and the average word length of the transcript (F-Score ~ Avg. Word Length: 0.479***). Either of the first two relationship with the age of the video or its duration are difficult to explain. They may allude to a slight tendency for creators to formalize their content as their channels grow over time, but this is speculation. The third relationship with average word length is a sensible one; more "formal" words, by the definition of Heylighen & Dewaele (1991), are typically more complex and therefore longer.

The second independent variable is the Readability Score, which primarily and logically correlates with the average word length (Readability Score ~ Avg. Word Length: 0.577***) of a transcript and the average sentence length (Readability Score ~ Avg. Sent Length: 0.624***). The Readability Score finds it origin in a list of "easy words" and most of those easy words are rather short. Thus, the longer the words on average are, the more "difficult" words will be observed, resulting in a higher Readability Score. The second relationship is also logical since the Average Sentence Length in words is an integral part of the readability equation.

The third and fourth independent variable, Lexical Density and MLCU, are both highly related empirically. Since most communication units contain one clause, and the MLCU

takes the number of words of those sentences it is bound to be highly related to a measure like Lexical Density which essentially reflects the prevalence of a mere subsection of the words within those clauses. Hence, their positive correlation (Lexical Density ~ MLCU: 0.962***) is also logically explained, yet theoretically irrelevant. The relationships of either variable with average sentence length and average word length can also be logically explained due to their empirical nature and do not form substantial findings.

Our last independent variable is Clausal Density and it does not correlate with many variables; however, it does correlate strongly positively with duration (Clausal Density ~ Duration: 0.827***). This finding is rather counter-intuitive, we would have expected a denser packaging of information to result in a need for less time to elaborate on said information. This does not seem to be the case, however, and additional research would be needed to understand this phenomenon in greater detail.

For our control variables, the main and most profound finding is the strong negative correlation between the age of the video and its duration. This finding can most probably be traced back to the monetization policy by YouTube. In recent years, videos longer than 10 minutes have been allowed more advertisement slots (Hutchinson, 2020). Secondly, optimizing watch time has grown as a greater priority for YouTube and this has been incorporated in the algorithm and reacted upon by creators (Sachs, 2017). This means, the more recent a video is, the longer it likely is to be.

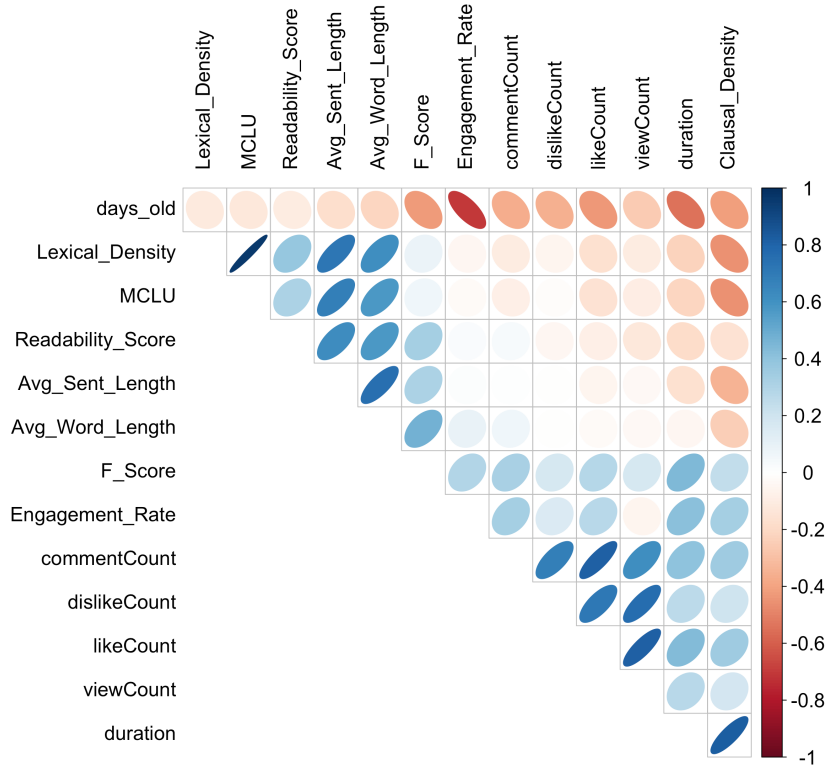


Figure 5.1 Correlogram

Table 5.2 Correlation Matrix

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. F-Score	-													
2. Readability Score	0.332***	-												
3. Lexical Density	0.085*	0.382***	-											
4. MLCU	0.064	0.314***	0.962***	-										
5. Clausal Density	0.249***	-0.152***	-0.457***	-0.460***	-									
6. Days old	-0.429***	-0.101**	-0.113**	-0.129***	-0.411***	-								
7. Duration	0.449***	-0.185***	-0.223***	-0.211***	0.827***	-0.544***	-							
8. Like Count	0.286***	-0.087*	-0.163***	-0.157***	0.350***	-0.438***	0.436***	-						
9. Dislike Count	0.174***	-0.043	-0.051	-0.014	0.205***	-0.351***	0.262***	0.710***	-					
10. Comment Count	0.328***	0.032	-0.101**	-0.086*	0.355***	-0.366***	0.403***	0.813***	0.683***	-				
11. Avg. Word Length	0.479***	0.577***	0.612***	0.577***	-0.247***	-0.211***	-0.049	-0.023	-0.002	0.061	-			
12. Avg. Sent. Length	0.313***	0.624***	0.722***	0.686***	-0.349***	-0.170***	-0.164***	-0.056	-0.006	0.009	0.760***	-		
13. View Count	0.177***	-0.126**	-0.101**	-0.094*	0.183***	-0.256***	0.277***	0.814***	0.767***	0.612***	-0.037	-0.035	-	
14. Engagement Rate	0.299***	0.021	-0.046	-0.028	0.340***	-0.709***	0.412***	0.278***	0.159***	0.330***	0.091*	0.012	-0.052	-

A Pair Plot is provided in appendix 9.2 to visualize the relationships and correlations between all the variables and provide an overview of their individual distributions.

5.1.3 Distribution of the Dependent Variables

Prior to investigating the hypothesized effects, we investigate the distribution of our dependent variables to aid our interpretation. Figures 5.1 and 5.2 show the density distribution of both our dependent variables. Figure 5.3 complements the former two plots by showing a violin representation of the distribution of each of the dependent variables and their dependencies. Appendix 9.3 contains a joint plot where we divide the distribution further and home in on the relationship between the two variables, albeit insignificant.

The density distribution plots visualize how especially the view counts for our sample are rather dispersed (take note of the logarithmic axes), the majority of view counts range between 1 million and 10 million. However, the sample also includes videos with less than 100.000 views and videos with more than 100 million views. The consequences of this large dispersion will be explored within the next section.

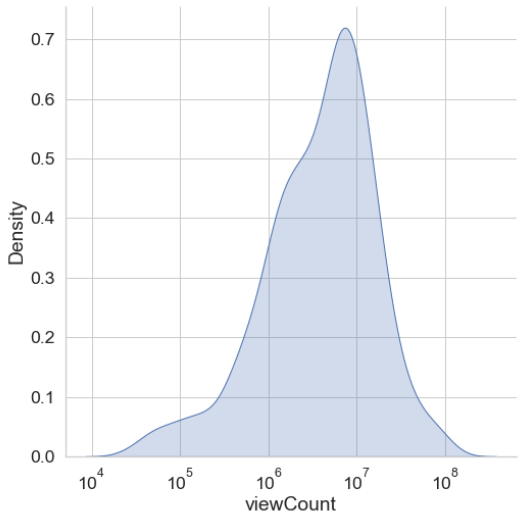


Figure 5.2 View Count Distribution (logarithmic x-axis)

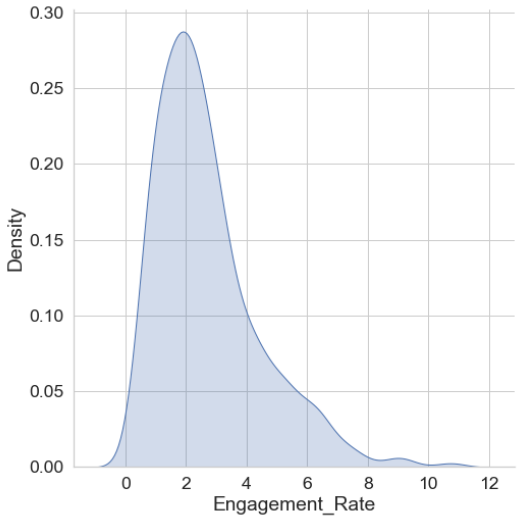


Figure 5.3 Engagement Rate Distribution

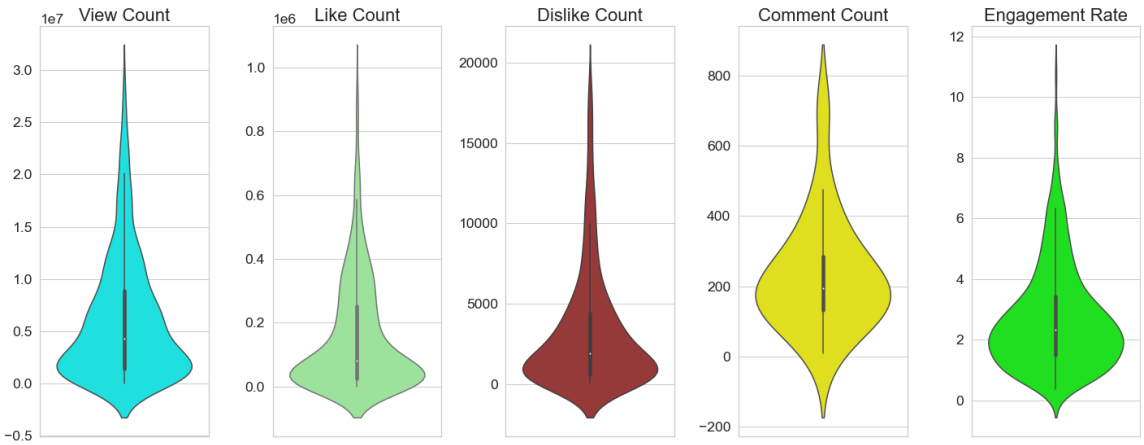


Figure 5.4 Violin Plots on Every Dependent Variable and Dependency

5.2 Regression Results

This section shall pertain to the results following the estimation of the empirical models as defined in section 4.4. As mentioned in the above section, a large dispersion of values exists within one of our dependent variables, View Count. The first subsection (5.2.1) will therefore elaborate on how this was navigated prior to us moving on with the regression results per each model variation (5.2.2 & 5.2.3).

5.2.1 View Count Variable Transformation

First of all, before the actual transformation we opted to divide each of the values by 1,000,000 to aid interpretation of the coefficients later on. Figure 5.5 depicts the results of transforming our variable. The graphs plot the fitted values (predictions) of our OLS models on the y-axis and the true (observed) values on the x-axis. The first graph, in blue, represents a model fitted on an untransformed dataset. Here we find that as the line approaches more of the outliers, the confidence interval becomes larger, and the predictions are consistently much lower than the true values. The model clearly becomes more inaccurate as it struggles to explain the variance found in the outliers. The graph in purple shows our model after the exclusion of outliers based on Tukey's fences (Tukey, 1977), which meant excluding any video that has more than 23,309,353 views. Excluding 41 outliers improved the model somewhat, with an increase of explanatory power from 0.131 to 0.212; however, even when excluding outliers as such, the values within our data set remain rather dispersed. To iron out this dispersion, the scale of our View Count variable is transformed to a logarithmic one. This improved the explanatory power to 0.266, more than double of what it originally was, and provides us a relatively steep line as depicted in the green graph. Other robustness checks have been included in appendix 9.4.

Figure 5.5 View Count Variable Transformation (based on model 5A)

5.2.2 OLS Models

In Table 5.3 we report the results of the first five empirical models estimated with OLS regression. The independent variables are introduced stepwise in an effort to gain a further understanding of the effects we observe. The models are divided over A variants and B variants which reflect the different dependent variables as defined in section 4.4. In the below section we will describe each of the variables per each model variant for which we observe significant effects. Each of the variables that were included within the regression had a variable inflation factor below 5 (9.4.3).

Opening with the models that estimate View Count, we find significant effects for three of our five measures of formality: the Lexical Density, Clausal Density and the F-Score. However, the effect of Lexical Density disappears once MLCU is introduced. This observation is likely motivated by the close relatedness of the two measures and can thus be ignored. Clausal Density is only tested in the final model however it is found significant (Clausal Density$_{5A}$: 0.011***). Since the dependent variable in this case is logarithmic, we ought to interpret the exponent of the beta coefficient, which is 1.011061 and transform this to a percentage. Therefore, one unit increase in clausal density leads to an increase of 1.1% of view counts. Within our sample, considering a mean of 8,481,122 views this leads to an average increase of views per unit increase Clausal Density of 93,292 views.

Perhaps the most substantial finding in the OLS regression estimates for View Count are the coefficients for the F-Score. Despite introducing more variables, the coefficients for the F-score remain statistically significant throughout all the models it is tested in (F-Score$_{1\text{-}5A}$: 0.006***), this leaves the finding very robust. Intuitively the coefficient entails that one standard deviation (≈ 70 units) increase in the F-Score leads to an increase of 3,541,208 views and a unit increase of 50,887 views on average.

The models for Engagement Rate do not find any significant results other than that of lexical density which we already explained in the passage above. We find that between the models the maximum adjusted R^2 we observe per dependent variable are 0.255 and 0.515 for View Count and Engagement Rate respectively. Considering the subject matter these are substantial numbers. Formality is likely only a very small determinant of success on YouTube; as touched upon in section 2.1, a plethora of scientific research is available on various other determinants of the success of YouTube content. However, whereas the residual standard error of the A models is quite low, that of the B models is large compared to the mean and standard deviation (RSE ≈ 1.200; $\mu = 2.71$; $\sigma = 1.72$), this alludes to a less-than-ideal model fit.

5.2.3 Fixed Effects Models

The second set of models are our fixed effects models which we report on in Table 5.4. These models are again divided in A and B variants and feature a stepwise introduction of independent variables. Where they differ from the OLS models is the fact that they find their basis on a different dataset. In the dataset used for models 6–10A&B we demean all of the variables by subtracting the group mean (based on channel name) from each value in the dataset. This controls for all constant unobserved effects that may be contained per channel.

The results reported in Table 5.4 include two main effects: those of the F-Score on both View Count as well as Engagement Rate. However, a third mildly significant ($p < 0.1$) effect is also contained that of clausal density (Clausal Density$_{10B}$: -0.007*). Intuitively interpreting estimates for the Engagement Rate is difficult due to the opaque nature of the variable; it is difficult to wrap one's head around what an Engagement Rate actually is. The best way to benchmark the finding in this case is to turn to the standard deviation of the dependent variable ($\mu = 2.71$; $\sigma = 1.72$). Consequently, we interpret the 0.007 decrease of the Engagement Rate with every unit increase of Clausal Density as being a decrease of merely 0.4% of a standard deviation, which leaves it negligible.

What is not negligible is the effects that the F-Score has remained to have on the View Count, albeit slightly reduced from its state in the OLS Models (F-Score$_{10A}$: 0.005***). However, when negating for channel related unobserved constants, one standard deviation increase in the F-Score remains to cause an influx in views of 2,951,006, and 42,406 views per unit increase. The F-Score also affects Engagement Rate significantly (F-Score$_{10B}$: -0.003***). Although at 0.2% of a standard deviation the effect is hardly remarkable, it is interesting to note that although the F-Score has a positive effect on views, the inverse might be true about engagement. This finding would be line with what Welbourne & Grant (2016) have found.

Finally, to offer a word on the quality of the models themselves, whereas the adjusted R^2 is acceptable for the OLS models, the fixed effects models see a vastly lower adjusted R^2 for those models that estimate View Count. This is in part due to the practice of demeaning the variables since the assumption then becomes that a demeaned value is equally capable in explaining the same amount of variance; however, this is not always the case. If we were to include dummy variables, the results would have been the same, but the explanatory power would have looked much greater (> 0.35). Supporting that theory is a relatively low residual standard error for the A models when compared to the mean and standard deviation of the outcome variable (RSE ≈ 1.536; $\mu = 8.481$; $\sigma = 12.857$).

Table 5.3 OLS Regression Results

	Dependent variable:									
	View Count in Millions (log)					Engagement Rate				
	(1_A)	(2_A)	(3_A)	(4_A)	(5_A)	(1_B)	(2_B)	(3_B)	(4_B)	(5_B)
F-Score	0.006***	0.006***	0.006***	0.006***	0.006***	0.001	0.0005	-0.0001	-0.0001	-0.0002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Readability Score		0.030	-0.002	-0.007	-0.069		0.027	-0.001	0.001	0.016
		(0.071)	(0.071)	(0.073)	(0.077)		(0.070)	(0.071)	(0.073)	(0.077)
Lexical Density			-0.005***	-0.004	-0.004			-0.005***	-0.005	-0.005
			(0.002)	(0.004)	(0.004)			(0.002)	(0.004)	(0.004)
MLCU				-0.001	0.0003				0.0002	0.0001
				(0.002)	(0.002)				(0.002)	(0.002)
Clausal Density					0.011***					-0.002
					(0.004)					(0.004)
Days old	-0.0001*	-0.0001*	-0.0001**	-0.0001**	-0.0001*	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Duration (in sec.)	0.0005**	0.001**	0.0004**	0.0004*	-0.0003	-0.0001	-0.0001	-0.0001	-0.0001	0.00003
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0004)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0004)
Avg. Word Length	0.370	0.361	0.561**	0.563**	0.613***	0.207	0.196	0.366	0.365	0.356
	(0.227)	(0.228)	(0.234)	(0.235)	(0.234)	(0.227)	(0.229)	(0.236)	(0.236)	(0.237)
Avg. Sent Length	0.003	0.002	0.018*	0.018*	0.021**	-0.030***	-0.031***	-0.017*	-0.017*	-0.018*
	(0.008)	(0.009)	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)
Constant	-1.245	-1.460	-1.925**	-1.881*	-1.690*	4.641***	4.452***	4.059***	4.040***	3.983***
	(0.827)	(0.977)	(0.980)	(0.993)	(0.991)	(0.827)	(0.962)	(0.968)	(0.980)	(0.986)
Observations	623	623	623	623	623	664	664	664	664	664
R^2	0.244	0.245	0.257	0.257	0.265	0.516	0.516	0.522	0.522	0.522
Adjusted R^2	0.238	0.237	0.249	0.248	0.255	0.512	0.512	0.516	0.516	0.515
Residual Std. Error	1.183 (df= 617)	1.184 (df= 616)	1.175 (df= 615)	1.176 (df= 614)	1.171 (df= 613)	1.203 (df= 658)	1.203 (df= 657)	1.197 (df= 656)	1.198 (df= 655)	1.199 (df= 654)
F Statistic	39.910*** (df = 5; 617)	33.243*** (df = 6; 616)	30.414*** (df = 7; 615)	26.583*** (df = 8; 614)	24.612*** (df = 9; 613)	140.318** * (df = 5; 658)	116.805** * (df = 6; 657)	102.176** * (df = 7; 656)	89.272*** (df = 8; 655)	79.310*** (df = 9; 654)

Note: *p**p***p<0.01

Table 5.4 Fixed Effects (Demeaned) OLS Regression Results

	Dependent variable:									
	View Count in Millions (log)					Engagement Rate				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
F-Score	0.005***	0.005***	0.005***	0.005***	0.005***	-0.003***	-0.002**	-0.002**	-0.002**	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Readability Score		0.122	0.099	0.108	0.087		-0.081	-0.097	-0.100	-0.067
		(0.097)	(0.098)	(0.101)	(0.105)		(0.069)	(0.070)	(0.072)	(0.074)
Lexical Density			-0.003	-0.006	-0.006			-0.002	-0.002	-0.002
			(0.002)	(0.006)	(0.006)			(0.002)	(0.004)	(0.004)
MLCU				0.001	0.001				-0.0003	-0.001
				(0.003)	(0.003)				(0.002)	(0.002)
Clauses					0.005					-0.007*
					(0.006)					(0.004)
Days old	-0.00002	-0.00003	-0.00004	-0.00004	-0.0001	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
duration	0.0002	0.0002	0.0001	0.0001	-0.0002	0.0001	0.0001	0.00004	0.00004	0.001
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0004)
Avg. Word Length	-0.274	-0.292	-0.182	-0.185	-0.149	0.056	0.071	0.144	0.145	0.092
	(0.299)	(0.300)	(0.309)	(0.310)	(0.313)	(0.214)	(0.214)	(0.221)	(0.221)	(0.223)
Avg. Sent. Length	-0.003	-0.007	0.001	0.0005	0.002	-0.016*	-0.013	-0.008	-0.008	-0.010
	(0.012)	(0.012)	(0.013)	(0.013)	(0.014)	(0.008)	(0.009)	(0.010)	(0.010)	(0.010)
Constant	2.040***	2.040***	2.040***	2.040***	2.040***	0.000	0.000	0.000	0.000	0.000
	(0.062)	(0.062)	(0.061)	(0.062)	(0.062)	(0.043)	(0.043)	(0.043)	(0.043)	(0.043)
Observations	623	623	623	623	623	664	664	664	664	664
R^2	0.042	0.044	0.048	0.048	0.049	0.514	0.515	0.516	0.516	0.518
Adjusted R^2	0.034	0.035	0.037	0.035	0.035	0.510	0.510	0.511	0.510	0.511
Residual Std. Error	1.536 (df = 617)	1.535 (df = 616)	1.534 (df = 615)	1.535 (df = 614)	1.536 (df = 613)	1.106 (df = 658)	1.106 (df = 657)	1.105 (df = 656)	1.106 (df = 655)	1.105 (df = 654)
F Statistic	5.404*** (df = 5; 617)	4.772*** (df = 6; 616)	4.383*** (df = 7; 615)	3.850*** (df = 8; 614)	3.482*** (df = 9; 613)	138.982*** (df = 5; 658)	116.122*** (df = 6; 657)	99.902*** (df = 7; 656)	87.290*** (df = 8; 655)	78.111*** (df = 9; 654)

Note: *p**p***p<0.01

6 DISCUSSION

The following section will offer further explanations, interpretations and justifications for the results of the entire study. We will do so by reflecting on the three main contributions this study set out to materialize as described in section 1.2. The first section (6.1) will give meaning to the results of the actual analysis and clarify how they can be used to further science communication. The second section (6.2) will home in on transcript analysis as a method and what merit it has. Lastly, the third section (6.3) will reflect on learnings drawn from the study related to formality as a construct for research. Limitations of the study are mentioned throughout the discussion where relevant but later also synthesized in the next chapter.

6.1 Dear Science Communicators,

In the early stages of this research a contrast was drawn between two schools of thought. On the one hand there was the study by Welbourne and Grant (2016) who found that creating a connection with your audience is a vital tool in furthering science communication on YouTube. However, the study by Irvine (1979) taught us that formality decreases how personable an interaction is perceived, which, in turn, lead us to believe formality would negatively affect the success of science communication on YouTube. On the other hand, there is the study by Riesch (2015) who claimed that informal communication may impede the transmission of scientific facts due to it perpetuating the perception of a so-called "in-group".

The results of this study lead us to side with the latter perspective. Within our analysis many of the variables found statistically insignificant effects (justified in later sections), except for the F-Score. However, the F-Score is the most robust measurement in the context of our study as it originated from well-established academic research in the topic of formality directly. All other measures were related indirectly. Hence, the F-Score was always a measure to watch with utmost attention and the fact that we observe it having a sizable positive and statistically significant effect means we ought to determine that formality does indeed seem to have a positive effect on the success of science communication on YouTube. The size of the effect is rather substantial at roughly 3 million additional views per every standard deviation increase and would warrant further research. Considering the aforementioned, it would be ill-advised to provided science communication content in too informal of a manner.

It should be noted, however, that we also observe a very small but negative effect on Engagement Rate. This finding, albeit of negligible size, would be more in line with the theory of Welbourne & Grant (2016) combined with that of Irvine (1979). It might be the case that

whereas formality benefits views on the short term, the impact of decreased engagement will be felt over the long term. Further research into the topic with a more varied sample would be needed to conclusively describe this observation.

The latter touches upon something very important; these results are far from infinitely generalizable to other domains. This sample only included five very successful channels on YouTube, each within a niche of science communication. Part of the reason for this is the highly restrictive nature of the YouTube API; viable sampling can only be realized based on channel ids. In addition, one must not confuse science communication with science education. The findings of a higher degree of formality increasing the popularity of science communication on YouTube cannot be held equivalent to a more successful transmission of scientific facts. Additional research would be required to explore the true influence of formality on a broader pool of content and to explore its role in facilitating the retention of the provided information (if existent at all).

6.2 Does the Promise of Transcript Analysis Live up to its Premise?

A second main contribution of the study was to explore transcript analysis since a thorough review of the literature did not return any previous instances of where the technique was applied for similar objectives. However, prior to discussing the merit of the technique a distinction needs to be made between three types of transcripts: computer generated, human generated and scripted.

The present study used computer generated transcripts which were based on audio tracks downloaded from selected videos. This complicated the data collection somewhat. For one, contemporary transcription software is impressively accurate but far from perfect. At time of writing this piece, the transcription accuracy lies around 80%. This means many nuances are either lost in translations (or rather transcription) or are replaced by incorrect phrases all together. These errors in the transcript directly impact the quality of the measures based on them. Less sophisticated measures such as the F-Score and the Readability score are not impacted as much; however, for measure such as Lexical Density, the MLCU and Clausal Density where it is required to parse a document for clauses it is very important to have exact and correct sentence structures. This may also explain why these measures almost exclusively returned insignificant results.

The latter two types of transcripts typically provide a much more accurate depiction of the utterances made by the speaker. In this case transcript analysis could certainly provide a strong tool for the research and extraction of new features that explain various facts on different types of audio-visual content that were hitherto overlooked. One suggestion for future research would be to work with creators to gain access to the scripts they have written themselves to base their content on.

However, science communication content on YouTube, and especially the popular content, is much more than just the words spoken in the videos. Transcript analysis will never pick up on the use of devices like illustrative examples or specific audio cues the background. Yet these exact devices may very well be the secret to why online video is such a popular medium for science communication and why the models featured a relatively low explanatory power. Future studies should certainly consider incorporating various features that are able to capture the manner of presentation in a more holistic sense, both audibly as well as visually.

6.3 The Computability of Formality

As mentioned profusely throughout this work; communicative formality is a diffuse and convoluted construct. This opaqueness makes formality difficult to measure. Doing so in a computational context adds another dimension of difficulty. As touched upon earlier, whereas measures such as the readability score or the F-Score are perfectly possible to define, more sophisticated measure such as Lexical Density, MLCU and Clausal Density are much more difficult to accurately define.

This difficulty in part stems from a lack of development of natural language processing tools. Natural Language Processing is a relatively young field (only truly having reached commercial viability in 2009) and collectively the various developers of tools have only scratched the surface in understanding the complexity of natural human language. Many of the methods currently available focus mostly on ways in which the greatest initial understanding of language can be won at scale. Constructs as formality, but also matters such as humour and sarcasm, reside in a deeper and more implicit expression of language. The application for understanding and being able to accurately define such constructs currently only caters to niche applications of Natural Language Processing. This means the conventional tools such as NLTK, spaCy or Stanza will sparsely suffice. Time constraints meant that within this study the development of custom solutions was out of the question; however, future studies should

consider becoming contributors to these libraries as there lies an opportunity in topics like this to make a tangible difference in the collective understanding of human language.

7 LIMITATIONS & RECOMMENDATIONS

Table 7.1 Synthesis of Limitations and Recommendations

	Limitations	*Recommendations*
1	The generalizability of the effects does not stretch beyond the very popular science communication channels.	Repeat the study at a larger scale by including more channels at more strongly varying levels of popularity.
2	This study focusses on popularity; however, that may only be part of the equation. This study is unable to make any claims on learning and information retention.	Perform a similar study in a domain related to education, specifically Science Education. It would be helpful to know what the role of formality is in furthering academic achievement and whether that is in line with what is popular.
3	Modern transcription software is far from perfect. Its use was required for this study, but it did hurt the accuracy of results.	In future work on this topic, one should use human transcribers or opt to work together with creators in an effort to gain access to their scripts.
4	What is spoken is only one element of what makes a video popular on YouTube.	If time constraints allow it, future researchers should make an effort to include as many features as possible that are able to encode the presentation more holistically.
5	Current Natural Language Processing tools do not pose as out-of-box solutions in defining some of the measures used in this study. Due to time constraints, the measures were not defined as accurately as possible.	Future researchers into this topic ought to expect to need to contribute to the modern libraries themselves in order to fit their needs of understanding language at a higher level of detail.

8 REFERENCES

Akinnaso, F. (1985). 'On the Similarities Between Spoken and Written Language', *Language and Speech*, pp. 323-359, 28(4). doi:10.1177/002383098502800401

Arthurs, J., Drakopoulou, S., Gandini, A. (2018). 'Researching YouTube', *Convergence*, pp. 3-15, 24(1). doi:10.1177/1354856517737222

Ashwell, D. (2016). 'The challenges of science journalism: The perspectives of scientists, science communication advisors and journalists from New Zealand', *Public Understanding of Science*, pp. 379-393, 25(3). doi:10.1177/0963662514556144

Barry, D., Marzouk, F., Chulak-Oglu, K., Bennett, D., Tierney, P., O'Keeffe, G. (2016). 'Anatomy education for the YouTube generation', *Anatomical Sciences Education*, pp. 90-96, 9(1). doi:10.1002/ase.1550

Bärtl, M. (2018). 'YouTube channels, uploads and views: A statistical analysis of the past 10 years', *Convergence*, pp. 16-32, 24(1). doi:10.1177/1354856517736979

Bello, R. (2005). 'Situational formality, personality, and avoidance-avoidance conflict as causes of interpersonal equivocation', *Southern Communication Journal*, pp. 285-299, 70(4). doi:10.1080/10417940509373335

Bird, S., Klein, E., Loper, E. (2009). *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., Stanford.

Bonney, R., Cooper, C., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K., Shirk, J. (2009). 'Citizen science: A developing tool for expanding science knowledge and scientific literacy', *BioScience*, pp. 977-984, 59(11). doi:10.1525/bio.2009.59.11.9

Brodersen, A., Scellato, S., Wattenhofer, M. (2012). 'YouTube Around the World: Geographic Popularity of Videos', *Proceedings of the 21st international conference on World Wide Web*, pp. 0-10.

Brossard, D. (2013). 'New media landscapes and the science information consumer', *Proceedings of the National Academy of Sciences of the United States of America*, pp. 14096-14101, 110(SUPPL. 3). doi:10.1073/pnas.1212744110

Burns, T., O'Connor, D., Stocklmayer, S. (2003). 'Science communication: A contemporary definition', *Public*

Understanding of Science, pp. 183-202, 12(2). doi:10.1177/09636625030122004

Campbell, K., Wright, K. (2002). 'On-line support groups: An investigation of relationships among source credibility, dimensions of relational communication, and perceptions of emotional support', *Communication Research Reports*, pp. 183-193, 19(2). doi:10.1080/08824090209384846

Carr, D., Oliver, M., Burn, A. (2010). 'Learning, Teaching and Ambiguity in Virtual Worlds', *Proceedings of Researching Learning in Virtual Environments International Conference*, pp. 83-93. doi:10.1007/978-1-84996-047-2_9

Chafe, W. L. (1982). 'Integration and involvement in speaking, writing, and oral literature', *In D. Tannen (Ed.), Spoken and written language: Exploring orality and literacy*. Norwood, NJ: Ablex.

Chatzopoulou, G., Sheng, C., Faloutsos, M. (2010). 'A first step towards understanding popularity in YouTube', *2010 INFOCOM IEEE Conference on Computer Communications Workshops*, doi:10.1109/INFCOMW.2010.5466701

Cheng, X., Liu, J., Dale, C. (2013). 'Understanding the characteristics of internet short video sharing: A youtube-based measurement study', *IEEE Transactions on Multimedia,* pp. 1184-1194, 15(5). doi:10.1109/TMM.2013.2265531

Cole, S., Cole, J. (1968). 'Visibility and the Structural Bases of Awareness of Scientific Research', *American Sociological Review*, pp. 397-413, 33(3).

Dale, E., Chall, J. (1949). 'The Concept of Readability'. *Readability*, pp. 1-47, 1

Deboer, G. (2000). 'Scientific Literacy: Another Look at Its Historical and Contemporary Meanings and Its Relationship to Science Education Reform', *Journal of Research in Science Teaching*, pp. 582-601, 37(6)

Dimopoulos. K., Koulaidis, V., Sklaveniti, S. (2003). 'Towards an Analysis of Visual Images in School Science Textbooks and Press Articles about Science and Technology', *Research in Science Education*, pp. 189-216, 33

Dobrian, F., Sekar, V., Awan, A., Stoica, I., Joseph, D., Ganjam, A., Zhan, J., Zhang, H. (2011). 'Understanding the Impact of Video Quality on User Engagement', *ACM SIGCOMM Computer Communication Review*, pp. 0-12, 41(4). doi:10.1145/2043164.2018478

Eisenhart, M., Finkel, E., Marion, S. (1996). 'Creating the Conditions for

Scientific Literacy: A Re-Examination', *American Educational Research Journal*, pp. 261-295, 33(2).

Figueiredo, F., Benevenuto, F., Almeida, J. (2011). 'The Tube over Time: Characterizing Popularity Growth of YouTube Videos', *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*.

Furnham, A. (1990). *Language and personality*. In H. Giles & W. P. Robinson (Eds.), *Handbook of language and social psychology* (p. 73–95). John Wiley & Sons.

Goobie, G. C., Guler, S. A., Johannson, K. A., Fisher, J. H., Ryerson, C. J. (2019). 'YouTube Videos as a Source of Misinformation on Idiopathic Pulmonary Fibrosis', AnnalsATS, pp. 572-579, 16(5).

Griffiths, R. (1992). 'Speech Rate and Listening Comprehension: Further Evidence of the Relationship', *TESOL Quarterly*, pp. 385-390, 26(2)

Halliday, M. A. K. (1985). *Spoken and Written Language*. 2nd Edition. Oxford University Press, Oxford.

Heylighen, F., Dewaele, J. (1999). *Formality of Language: definition, measurement and behavioral determinants*. Free University of Brussels, Brussels

Heylighen, F. (1999). 'Advantages and limitations of formal expression', *Foundations of Science*, pp. 25-56, 4. doi:10.1023/A:1009686703349

Heylighen, F., Dewaele, J. (1999). 'Variation in the Contextuality of Language: An Empirical Measure', *Foundations of Science*, pp. 293-340, 7

Horowitz, M., Newman, J. (1964). 'Spoken and written expression: An experimental analysis', *The Journal of Abnormal and Social Psychology*, pp. 640-647, 68(6). doi:10.1037/h0048589

Hutchinson, A. (2020). YouTube's Adding More Ads, with Mid-Roll Breaks Available in Shorter Videos. Accessed on the 20th of July, 2021 at https://www.socialmediatoday.com/news/youtubes-adding-more-ads-with-mid-roll-breaks-available-in-shorter-videos/581173/

Hussin, M., Frazier, S., Thompson, J. (2011). 'Fat stigmatization on YouTube: A content analysis', *Body Image*, pp. 90-92, 8(1). doi:10.1016/j.bodyim.2010.10.003

Irvine, J. (1979). 'Formality and Informality in Communicative Events', *American Anthropologist*, pp. 773-790, 81(4). doi:10.1525/aa.1979.81.4.02a00020

Jaffar, A. (2012). YouTube: An emerging tool in anatomy education, Anatomical Sciences Education, pp.

158-164, 5(3). doi:10.1002/ase.1268

Kawamoto, S., Nakayama, M., Saijo, M. (2013). 'A survey of scientific literacy to provide a foundation for designing science communication in Japan', *Public Understanding of Science*, pp. 674-690, 22(6). doi:10.1177/0963662511418893

Keelan, J., Pavri-Garcia, V., Tomlinson, G., Wilson, K. (2007). 'YouTube as a Source of Information on Immunization: A Content Analysis', *Journal of the American Medical Association*, pp. 2482-2482, 298(21)

Kim, J. (2012). 'The institutionalization of youtube: From user-generated content to professionally generated content', *Media, Culture and Society*, pp. 53-67, 34(1). doi:10.1177/0163443711427199

Kraut, R., Resnick, P. (2012). 'Encouraging contribution to online communities, Building Successful Online Communities', pp. 21-77, 2

Lamb, C. T., Gilbert, S. L., Ford, T. A. (2018). 'Tweet success? Scientific communication correlates with increased citations in Ecology and Conservation', PeerJ 6:e4564. doi: 10.7717/peerj.4564

Laugksch, R. (1999). 'Scientific Literacy: A Conceptual Overview', *Science Education*, pp. 71-94, 84. doi:10.1002/(SICI)1098-237X(200001)84:13.0.CO;2-C

Matias, A., Dias, A., Gonçalves, C., Vicente, P. N. and Mena, A. L. (2020). 'Science communication for social inclusion: exploring science & art approaches'. *JCOM 20* (02), A05. doi:10.22323/2.20020205

Mehl, M., Robbins, M., Holleran, S. (2012). 'How Taking a Word for a Word Can Be Problematic: Context-Dependent Linguistic Markers of Extraversion and Neuroticism', *Journal of Methods and Measurement in the Social Sciences*, pp. 30-50, 3(2)

Miller, J. (1998). 'The measurement of civic scientific literacy', Public Understand of Science, pp. 203-223, 7

Newman, E., Schwarz, N. (2018). 'Good Sound, Good Research: How Audio Quality Influences Perceptions of the Research and Researcher', *Science Communication*, pp. 246-257, 40(2). doi:10.1177/1075547018759345

Nippold, M., Cramond, P., Hayward-Mayhew, C. (2014). 'Spoken language production in adults: Examining age-related differences in syntactic complexity'. *Clinical Linguistics and Phonetics*, pp. 195-207, 28(3).

doi:10.3109/02699206.2013.841292

Nippold, M., Frantz-Kaspar, M., Vigeland, L. (2017). 'Spoken language production in young adults: Examining syntactic complexity', *Journal of Speech, Language, and Hearing Research*, pp. 1339-1347, 60(5). doi:10.1044/2016_JSLHR-L-16-0124

Norris, S., Phillips, L. (2003). 'How Literacy in Its Fundamental Sense Is Central to Scientific Literacy', *Science Education*, pp. 224-240, 87(2). doi:10.1002/sce.10066

Paek, H., Kim, K., Hove, T. (2010). 'Content analysis of antismoking videos on YouTube: Message sensation value, message appeals, and their relationships with viewer responses', *Health Education Research*, pp. 1085-1099, 25(6) doi:10.1093/her/cyq063

Pavlick, E., Tetreault, J. (2016). 'An Empirical Analysis of Formality in Online Communication', *Transactions of the Association for Computational Linguistics*, pp. 61-74, 4. doi:10.1162/tacl_a_00083

Peña, J., Blackburn, K. (2013). 'The Priming Effects of Virtual Environments on Interpersonal Perceptions and Behaviors'. Journal of Communication, pp. 703-720, 63(4). doi:10.1111/jcom.12043

Pinto, H., Almeida, J., Gonçalves, M. (2013). 'Using early view patterns to predict the popularity of YouTube videos', *WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pp. 365-374. doi:10.1145/2433396.2433443

Richards, J., Platt, J., Weber, H., Inman, P., Inman, P. (1986). 'Longman Dictionary of Applied Linguistics', *RELC Journal*, 17(2). doi:10.1177/003368828601700208

Riesch, H. (2015). 'Why did the proton cross the road? Humour and science communication', *Public Understanding of Science*, pp. 768-775, 24(7). doi:10.1177/0963662514546299

Ryder, J. (2001). 'Identifying science understanding for functional scientific literacy', *Studies in Science Education*, pp. 1-44, 36(1). doi:10.1080/03057260108560166

Sachs, E. (2017). 4 Ways to Increase YouTube Watch Time. Accessed on the 20th of July at https://www.socialmediaexaminer.com/4-ways-to-increase-youtube-watch-time/

Scheufele, D., Krause, N. (2019). 'Science audiences, misinformation, and fake

news', *Proceedings of the National Academy of Sciences of the United States of America*, pp. 7662-7669, 116(16). doi:10.1073/pnas.1805871115

Schütz, A. (1946). 'THE WELL-INFORMED CITIZEN: An Essay on the Social Distribution of Knowledge', *Social Research*, pp. 463-478, 13(4)

Science Magazine. (2021). The science stories likely to make headlines in 2021. (Online) Retrieved from: https://www.sciencemag.org/news/2020/12/science-stories-likely-make-headlines-2021

Scott, B. (2021). The Dale-Chall 3,000 Word List for Readability Formulas. *Readability Formulas.* Accessed on the 1st of July, 2021 at https://readabilityformulas.com/articles/dale-chall-readability-word-list.php.

Stellefson, M., Chaney, B., Ochipa, K., Chaney, D., Haider, Z., Hanik, B., Chavarria, E., Bernhardt, J. (2014). 'YouTube as a source of chronic obstructive pulmonary disease patient education: A social media content analysis', Chronic Respiratory Disease, pp. 61-71, 11(2). doi:10.1177/1479972314525058

Tauroza, S., Allison, D. (1990). 'Speech Rates in British English', Applied Linguistics, pp. 90-105, 11(1). Doi: 11.1.90/255991

Tukey, J. (1997). *Exploratory data analysis*. 1st Edition. Addison-Wesley Pub. Co., Reading.

Valenti, J. (1999). 'Commentary: How Well Do Scientists Communicate to Media', *Science Communication*, pp. 172-178, 21(2)

Welbourne, D., Grant, W. (2016). 'Science communication on YouTube: Factors that affect channel and video popularity', *Public Understanding of Science*, pp. 706-718, 25(6). doi:10.1177/0963662515572068

Youtube. (n.d.a). YouTube About. (Online) Retrieved from: https://www.youtube.com/intl/nl/about/

Youtube. (n.d.b). API Reference. (Online) Retrieved from: https://developers.google.com/youtube/v3/docs

Yu, H., Xie, L., Sanner, S. (2015). 'The Lifecyle of a Youtube Video: Phases, Content and Popularity', *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pp. 533-542

9 APPENDIX

9.1 Appendix I: Data Collection Procedures

9.1.1 YouTube API: .search()

9.1.1.1 Configuration:

Table 9.1 Video Search Criteria

Parameter	Description	Configured values
`part`	Refers to the type of datapoints you desire to retrieve.	Snippet
`type`	Type of subject you are after investigating (channel or video)	Video
`maxResults`	Number of results you allow the API to compile	600
`channelId`	Unique identifier for the channel you aim to investigate videos of	Different channel ids for the 5 channels, please be referred to the code example below.
`videoDuration`	Filters video search results based on their duration.	Medium – Only include videos that are between four and 20 minutes long (inclusive)
`pageToken`	Every query retrieves a maximum of 50 results, to get the next 50 results of your search you need to query again and pass the next page token from the last response.	nextPageToken

9.1.1.2 Code Example:

```
# Allocate credentials:
from googleapiclient.discovery import build

# Api Keys
api_key = "####"

# Session Build
youtube = build('youtube', 'v3', developerKey = api_key)
```

```
# Variant 1 Sample Generation
channelids = {'Veritasium'  : 'UCHnyfMqiRRG1u-2MsSQLbXA',
              'VSauce'      : 'UC6nSFpj9HTCZ5t-N3Rm3-HA',
              'Kurzgesagt'  : 'UCsXVk37bltHxD1rDPwtNM8Q',
              'Mark Rober'  : 'UCY1kMZp36IQSyNx_9h4mpCg',
              'asapSCIENCE' : 'UCC552Sd-3nyi_tk2BudLUzA'}

n     = 600
iter  = range(1, (int(n/50+1)))
order = 'rating'

Raw_sample_V1 = pd.DataFrame(columns = ["Video.ID", "Title", "Channel_Name"])

#Iterate through Channels
for channelid in channelids.items():

    print(channelid[0])

    #Iterate iter number of times to fulfill n; there is a maximum of 50 results per search.
    for i in iter:

        if i == 1:

            # Search Request
            request = youtube.search().list(
                part       ="snippet",
                type       = "video",
                maxResults = n,
                channelId  = channelid[1],
            )

            # Save response
            response = request.execute()

            # Unpack Respons
            rows = []

            for item in response['items']:

                    rows.append([item['id']['videoId'],
                                 item['snippet']['title'],
                                 item['snippet']['channelTitle']])

            video_sample = pd.DataFrame(rows, columns = ["Video.ID", "Title", "Channel_Name"])
            print(f'{len(video_sample)} out of {n}')

        else:
            try:
                # Search Request
                request = youtube.search().list(
                    part       = "snippet",
                    type       = "video",
                    maxResults = n,
                    channelId  = channelid[1],
                    pageToken  = response['nextPageToken']
                )

                # Save response
                response = request.execute()

                # Unpack Respons
                rows = []

                for item in response['items']:

                    rows.append([item['id']['videoId'],
                                 item['snippet']['title'],
                                 item['snippet']['channelTitle']])

                video_sample_temp = pd.DataFrame(rows, columns = ["Video.ID", "Title", "Channel_Name"])
                video_sample = video_sample.append(video_sample_temp)
                print(f'{len(video_sample)} out of {n}')

            except(KeyError):
                print("Results exhausted")
                break

    #Cleaning:
    to_delete = ['#short',
                 ' prank']
    video_sample = video_sample[~video_sample['Title'].str.contains('|'.join(to_delete))]

    #Sampling:
    if len(video_sample) > 200:
        sample = video_sample.sample(n=200,
                                     random_state=123,
                                     replace = True)
    else:
        sample = video_sample

    Raw_sample_V1 = Raw_sample_V1.append(sample)
    print(f'The {channelid[0]} sample has been saved! \n')

#Output:
Raw_sample_V1.to_csv(f'/Users/Gerbrand/Documents/EUR RSM/Master/Thesis/Sample V1/Raw_Sample_V1.csv',
                     sep=';',
                     index=False,
                     encoding='utf-8')
```

9.1.2 YouTube API: .videos() Configuration

9.1.2.1 Configuration:

Table 9.2 Statistics Search Criteria

Parameter	Description	Configured values
`part`	Refers to the type of datapoints you desire to retrieve.	snippet, statistics, contentDetails
`id`	To retrieve statistics, you have to go off a video-to-video basis; the id refers to the unique identifier of the video you aim to investigate.	Video

9.1.2.2 Code Example:

```
def RetrieveStats(df):

    #Initialize dictionary for the data points to collect:
    stats = {"publishedAt"   : [],
             "duration"      : [],
             "definition"    : [],
             "viewCount"     : [],
             "likeCount"     : [],
             "dislikeCount"  : [],
             "favoriteCount" : [],
             "commentCount"  : []}

    # Execute request per Video ID
    for i in range(0,len(df)):

        vid = df.iloc[i]['Video.ID']

        # Formalize Request:
        request = youtube.videos().list(
                part ="snippet,statistics,contentDetails",
                id   = vid
            )

        # Save Response
        response = request.execute()

        # Store the data in the dictionary
        try:
            stats['publishedAt'].append(response['items'][0]['snippet']['publishedAt'])
        except(KeyError):
            stats['publishedAt'].append(np.nan)
        try:
            stats['duration'].append(response['items'][0]['contentDetails']['duration'])
        except(KeyError):
            stats['duration'].append(np.nan)
        try:
            stats['definition'].append(response['items'][0]['contentDetails']['definition'])
        except(KeyError):
            stats['definition'].append(np.nan)
        try:
            stats['viewCount'].append(response['items'][0]['statistics']['viewCount'])
        except(KeyError):
            stats['viewCount'].append(np.nan)
        try:
            stats['likeCount'].append(response['items'][0]['statistics']['likeCount'])
        except(KeyError):
            stats['likeCount'].append(np.nan)
        try:
            stats['dislikeCount'].append(response['items'][0]['statistics']['dislikeCount'])
        except(KeyError):
            stats['dislikeCount'].append(np.nan)
        try:
            stats['favoriteCount'].append(response['items'][0]['statistics']['favoriteCount'])
        except(KeyError):
            stats['favoriteCount'].append(np.nan)
        try:
            stats['commentCount'].append(response['items'][0]['statistics']['commentCount'])
        except(KeyError):
            stats['commentCount'].append(np.nan)

        # progress report:
        if i % 50 == 0:
            print(f'We are {(i/len(df)*100):.1f}% of the way there!')

    # Store data as a dataframe and concatenate it to the original:
    stats = pd.DataFrame(stats)
    df = pd.concat([df, stats], axis = 1)

    # Response summary:
    print(f"\nWe couldn't find at least 1 statistic for {df.isna().sum().max()} videos. \nSee a data loss overview belo
    print(df.isna().sum())

    return df
```

Response:

```
# Read data
raw_sample_V1 = pd.read_csv("/Users/Gerbrand/Documents/EUR RSM/Master/Thesis/Sample V1/Raw_Sample_V1.csv", ';')
#raw_sample_V2 = pd.read_csv("/Users/Gerbrand/Documents/EUR RSM/Master/Thesis/Sample V2/Raw_Sample_V2.csv", ';')

raw_sample_V1 = RetrieveStats(raw_sample_V1)
```

```
We are 0.0% of the way there!
We are 6.0% of the way there!
We are 11.9% of the way there!
We are 17.9% of the way there!
We are 23.9% of the way there!

We are 29.8% of the way there!

We are 35.8% of the way there!

We are 41.8% of the way there!

We are 47.7% of the way there!
We are 53.7% of the way there!
We are 59.7% of the way there!
We are 65.6% of the way there!
We are 71.6% of the way there!
We are 77.6% of the way there!
We are 83.5% of the way there!

We are 89.5% of the way there!
We are 95.5% of the way there!


 We couldn't find at least 1 statistic for 12 videos.
 See a data loss overview below:

Video.ID           0
Title              0
Channel_Name       0
publishedAt        0
duration           0
definition         0
viewCount         12
likeCount         12
dislikeCount      12
favoriteCount      0
commentCount      12
dtype: int64
```

9.1.3 youtube-dl Configuration

9.1.3.1 Configuration:

Table 9.3 youtube-dl Configuration

Command:	
"youtube-dl -f 251 http://www.youtube.com/watch?v={i} -o /Volumes/Samsung_T5/Thesis/Audio/V1/{i}.webm'"	
Break down:	
Youtube-dl	Call the library
-f 251	Only retrieve the webm track
URL	Video URL with id between brackets
-o filepath	Specify the directory where to store the audio file

9.1.3.2 Code Example:

```
# To download the audio files from youtube videos we'll be using the Youtube-dl library
# This library works on the command line, however we can use subprocess to control a shell
# at the command line with python through Jupyter notebooks.
import subprocess
import glob # To save time, we try to exclude videos we already downloaded.
downloaded = glob.glob(f"/Volumes/Samsung_T5/Thesis/Audio/{variant}/*.webm")
downloaded = [item.replace(f"/Volumes/Samsung_T5/Thesis/Audio/{variant}/", "") for item in downloaded]
downloaded = [item.replace(".webm", "") for item in downloaded]

to_dl = Sample_V1[~Sample_V1['Video.ID'].isin(downloaded)]
to_dl.reset_index(drop=True,inplace=True)

for i in to_dl['Video.ID']:
    command = f'youtube-dl -f 251 http://www.youtube.com/watch?v={i} -o /Volumes/Samsung_T5/Thesis/Audio/V1/{i}.webm'
    subprocess.call(command, shell = True)

    # Progress report
    current = to_dl[to_dl['Video.ID'] == i].index[0] + 1
    every   = len(to_dl['Video.ID'])

    print(f'{current} out of {every} audio files has been downloaded.')
    print(f'We are {((current/every)*100):.1f}% of the way there! \n')
```

```
1 out of 428 audio files has been downloaded.
We are 0.233644859813084O8% of the way there!

4 out of 428 audio files has been downloaded.
We are 0.9345794392523363% of the way there!

6 out of 428 audio files has been downloaded.
We are 1.4018691588785046% of the way there!

10 out of 428 audio files has been downloaded.
We are 2.3364485981308 41% of the way there!

11 out of 428 audio files has been downloaded.
We are 2.570093457943925% of the way there!

13 out of 428 audio files has been downloaded.
We are 3.0373831775700935% of the way there!
```

9.1.4 IBM Watson Speech-to-Text Service Configuration

9.1.4.1 Configuration:

```
from ibm_watson import SpeechToTextV1
from ibm_watson.websocket import RecognizeCallback, AudioSource
from ibm_cloud_sdk_core.authenticators import IAMAuthenticator
# Setup credentials
credentials = {
    'user':('uwM84UUngydIEgHU0gzmvVkRkIA_ZvTXj##########',
            'https://api.eu-gb.speech-to-text.watson.cloud.ibm.com/instances/373264b1-d978-4edd-ad4e-############'),
}
# Setup service
authenticator = IAMAuthenticator(credentials['kian'][0])
stt = SpeechToTextV1(authenticator = authenticator)
stt.set_service_url(credentials['kian'][1])
```

9.1.4.2 Code Example:

```
def Transcribe(df, variant='V1'):

    # First up, let's make sure that all of the Video ID's we pass have actually also
    # been succesfully downloaded; this is in an effort to reduce errors.
    import glob
    downloaded = glob.glob(f"/Volumes/Samsung_T5/Thesis/Audio/{variant}/*.webm")
    downloaded = [item.replace(f"/Volumes/Samsung_T5/Thesis/Audio/{variant}/", "") for item in downloaded]
    downloaded = [item.replace(".webm", "") for item in downloaded]

    df = df[df['Video.ID'].isin(downloaded)]

    # Since the API limits the number of calls we can give, it is important to conserve
    # our calling capacity. To do so, we make sure to only include videos that are yet
    # to be transcribed. We include this code in our loop so it redfines at every iteration
    # to help us track progress.
    transcribed = glob.glob(f"/Volumes/Samsung_T5/Thesis/Text/{variant}/*.txt")
    transcribed = [item.replace(f"/Volumes/Samsung_T5/Thesis/Text/{variant}/", "") for item in transcribed]
    transcribed = [item.replace(".txt", "") for item in transcribed]

    transcribed.append('b0cakXwiZ8s')
    transcribed.append('GeyDf4ooPdo') # Troublemakers
    transcribed.append('iqKdEhx-dD4')
    transcribed.append('5J3peD8LZ5o')
    transcribed.append('tH2tKigOPBU')
    transcribed.append('P_6my53IlxY')
    transcribed.append('tMKXbLBgkEc')
    transcribed.append('PCKogFDM3Zg')

    df = df[~df['Video.ID'].isin(transcribed)]

    # Second, Let's get transcribing!

    # Initialize an empty dictionary to store results:
    res = {}

    for i in df['Video.ID']:

        print(i)

        with open(f'/Volumes/Samsung_T5/Thesis/Audio/{variant}/{i}.webm', 'rb') as f:
            res[f'{i}'] = stt.recognize(audio=f,
                                        content_type='audio/webm',
                                        model='en-US_NarrowbandModel',
                                        continuous=True).get_result()

    # Third, we ought to process the results into .txt documents
        try:
            results = res[f'{i}']['results']

            text = [result['alternatives'][0]['transcript'].rstrip() + '.\n' for result in res[i]['results']]
            text = [para[0].title() + para[1:] for para in text]

            transcript = ''.join(text)

            with open(f'/Volumes/Samsung_T5/Thesis/Text/V1/{i}.txt', 'w') as out:
                out.writelines(transcript)

        except(KeyError):
            pass

        # Progress report
        transcribed = glob.glob(f"/Volumes/Samsung_T5/Thesis/Text/{variant}/*.txt")
        transcribed = [item.replace(f"/Volumes/Samsung_T5/Thesis/Text/{variant}/", "") for item in transcribed]
        transcribed = [item.replace(".txt", "") for item in transcribed]

        current = len(transcribed) + 1
        every   = len(downloaded)

        print(f'{current} out of {every} files has been transcribed.')
        print(f'We are {((current/every)*100):.1f}% of the way there! \n')

    return
```

9.2 Appendix II: Pair Plot

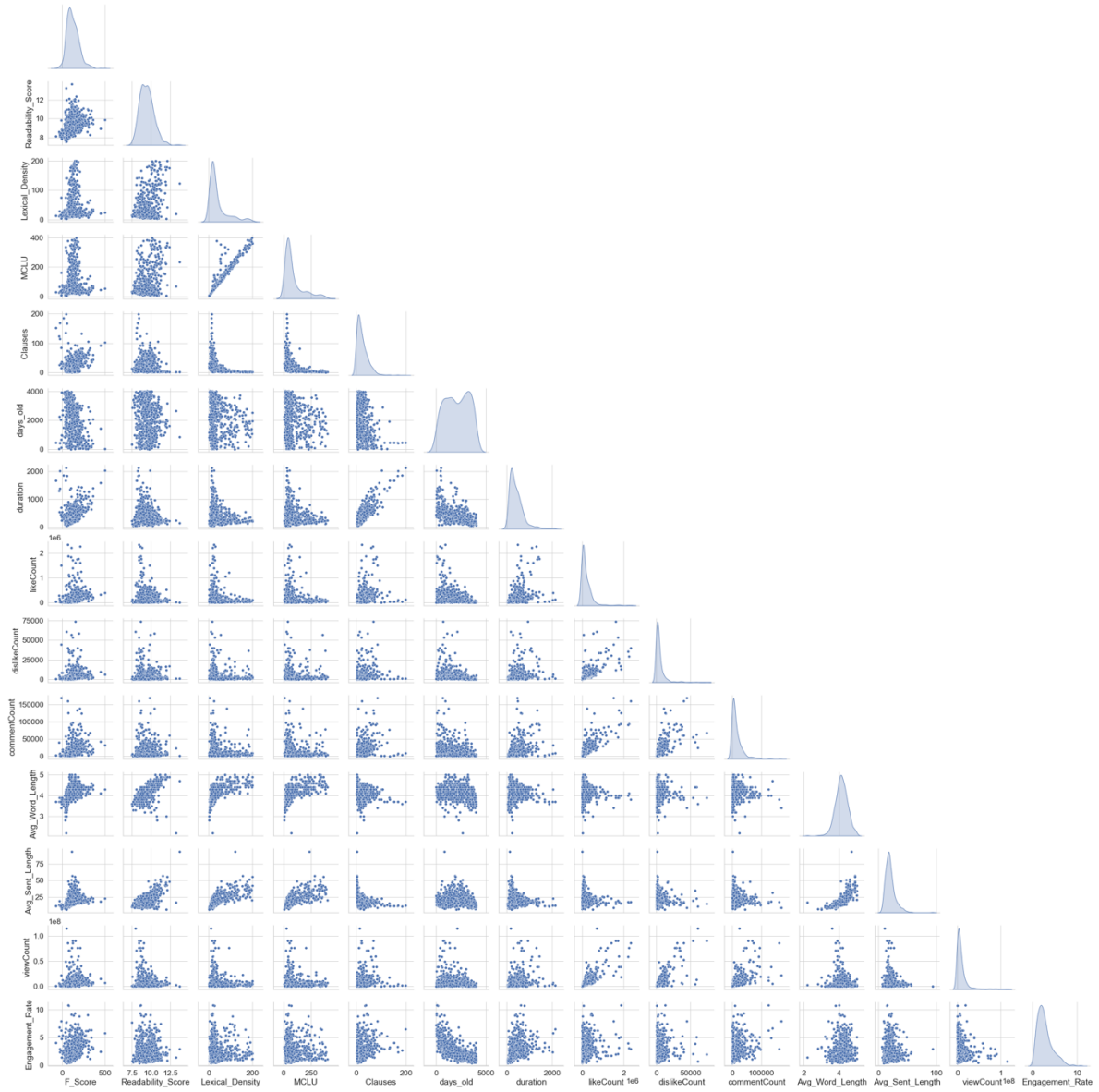


Figure 9.1 Pair Plot

9.3 Appendix III: Joint Plot: Dependent Variables

Figure 7.1 shows the distributions of both dependent variables split over each of the investigated channels. In the middle it shows a scatter plot where the variables show an insignificant correlation of -0.052. What can be distilled from this plot is that many of Mark Rober's videos have rather high view counts (this plot is limited to showing 60,000,000 views). However, in the distribution we can see that he makes very few videos. In terms of engagement, it looks like Kurzgesagt is doing rather well. It seems as though they have the best balance in terms of views and engagement out of all the channels. The least engaging videos were by Vsauce, and they score middle-off-road in terms of view count. However, Vsauce is the oldest and largest channel within the sample and suffered the most from data loss.

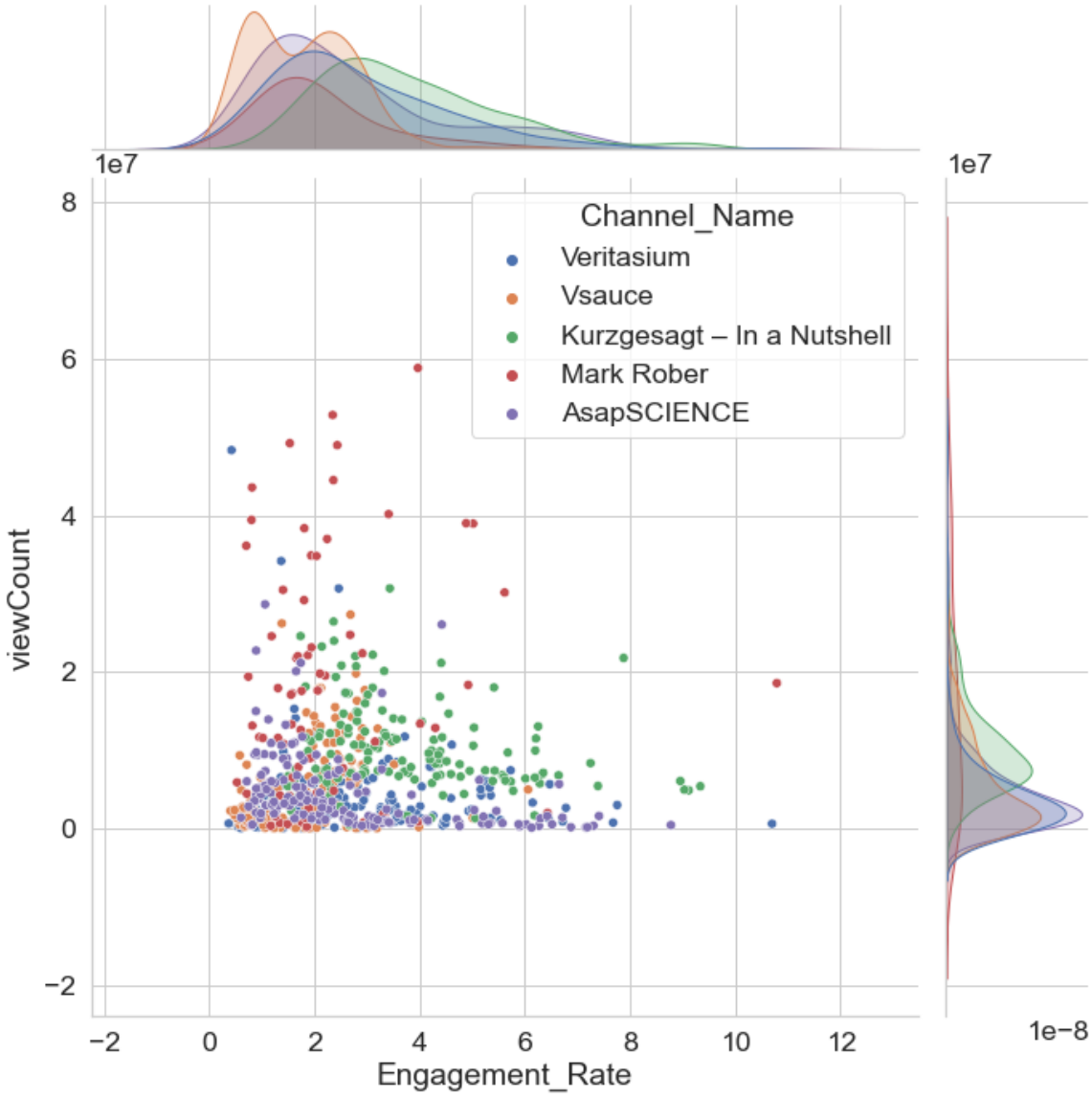


Figure 9.2 Joint Plot of the Dependent Variables

9.4 Appendix IV: Robustness Checks

Significant credit is due to Tirthajyoti Sarkar for his script that was used to create many of the plots below.

9.4.1 Fitted vs. True Plot Engagement Rate

After having discussed the plots for View Count in section 5.2.1, it would be good to also discuss the same plot for Engagement Rate. This is contained here within the appendix, however, since the results look perfectly workable and do not require any transformation.

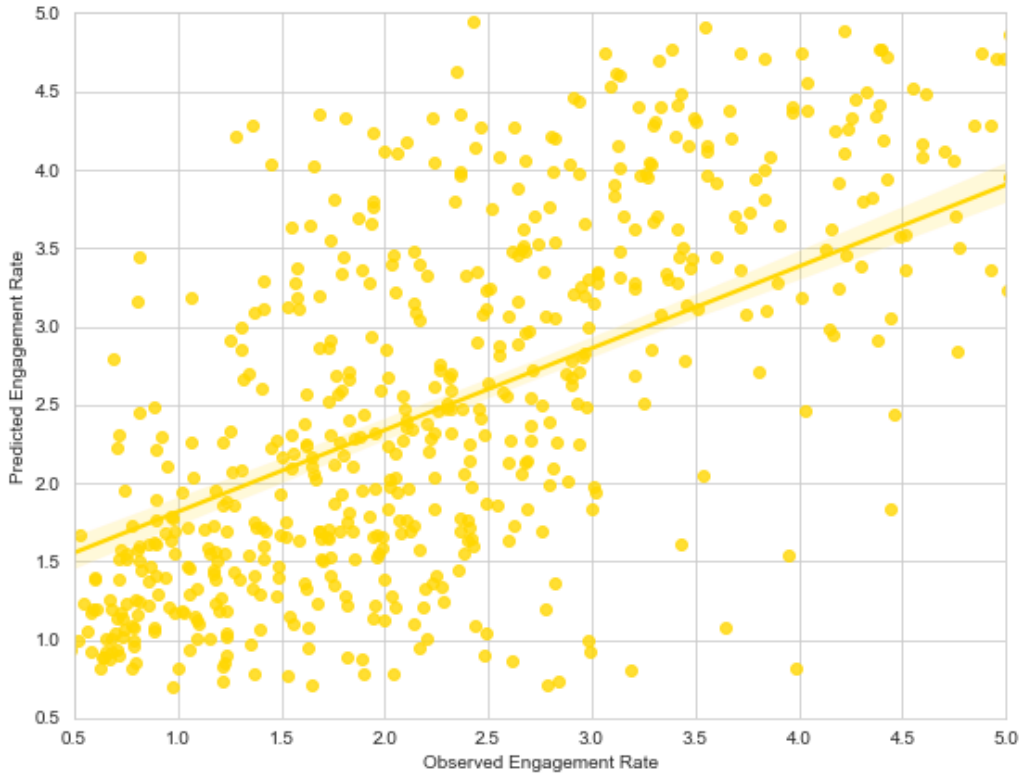


Figure 9.3 Fitted vs. Observed Plot Engagement Rate

9.4.2 Non-Linearity

The non-linearity criterium seems to be met for the most part. When looking at the Pair Plot in section 9.2 we find that only between Days old and Engagement Rate a non-linear relationship seems to exist. This being a control variable that is of no concern.

9.4.3 Multicollinearity

When calculating the variable inflation factor (ViF) for all the variables within the dataset we find that 3 variables ought to be excluded. After exclusion of Character Count, Word Count and Sentence Count, we are left with a table that only contain 2 variables that truly do not meet the norm for multicollinearity, and that are MLCU and Lexical Density. However, as we explain in chapter 5, this occurrence can be explained logically and therefore does not motivate exclusion of the variables even though they do not benefit the quality of the model itself.

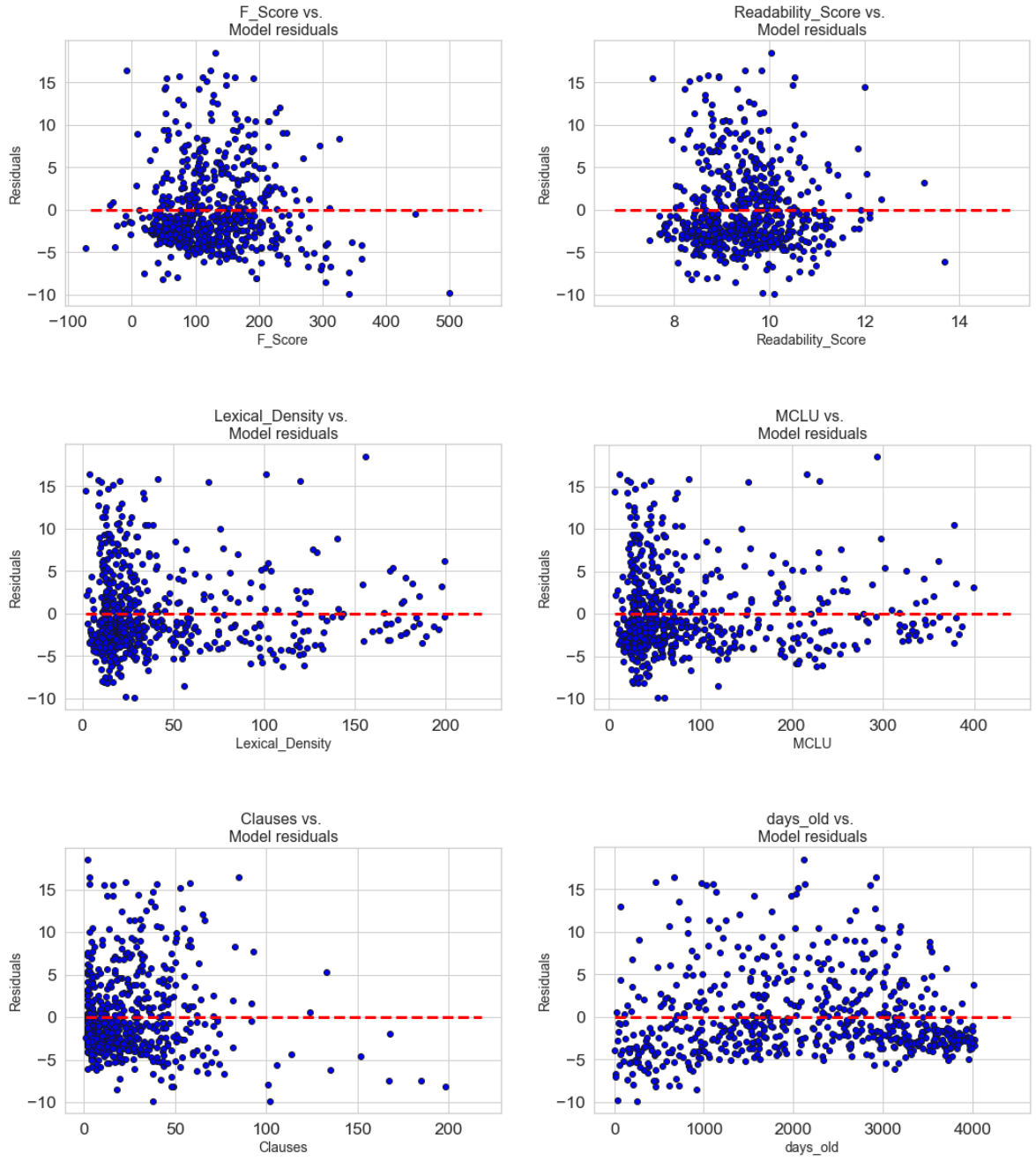
Table 9.4 Multicollinearity

Full Dataset		*Post Exclusion*	
Features	ViF	Features	ViF
F-Score	4.78	F-Score	2.16
Readability Score	2.81	Readability Score	2.21
Lexical Density	**15.55**	Lexical Density	**15.31**
MLCU	**14.70**	MLCU	**14.46**
Clausal Density	9.13	Clausal Density	5.12
Days Old	1.80	Days Old	1.64
Duration	**40.52**	Duration	5.54
Word Count	122013	Avg. Word Length	3.10
Character Count	1116.09	Avg. Sent. Length	3.77
Sentence Count	**34.74**		
Avg. Word Length	4.43		
Avg. Sent. Length	4.90		

9.4.4 Independence of Residuals

9.4.4.1 View Count

Below in Figure 9.4 we find the residual plots for each of the predictor variables that estimate view count. Aside from some distribution asymmetry and outliers we don't see many particular patterns that violate the independence of residuals assumption. Only days old seems to show a slight tendency.



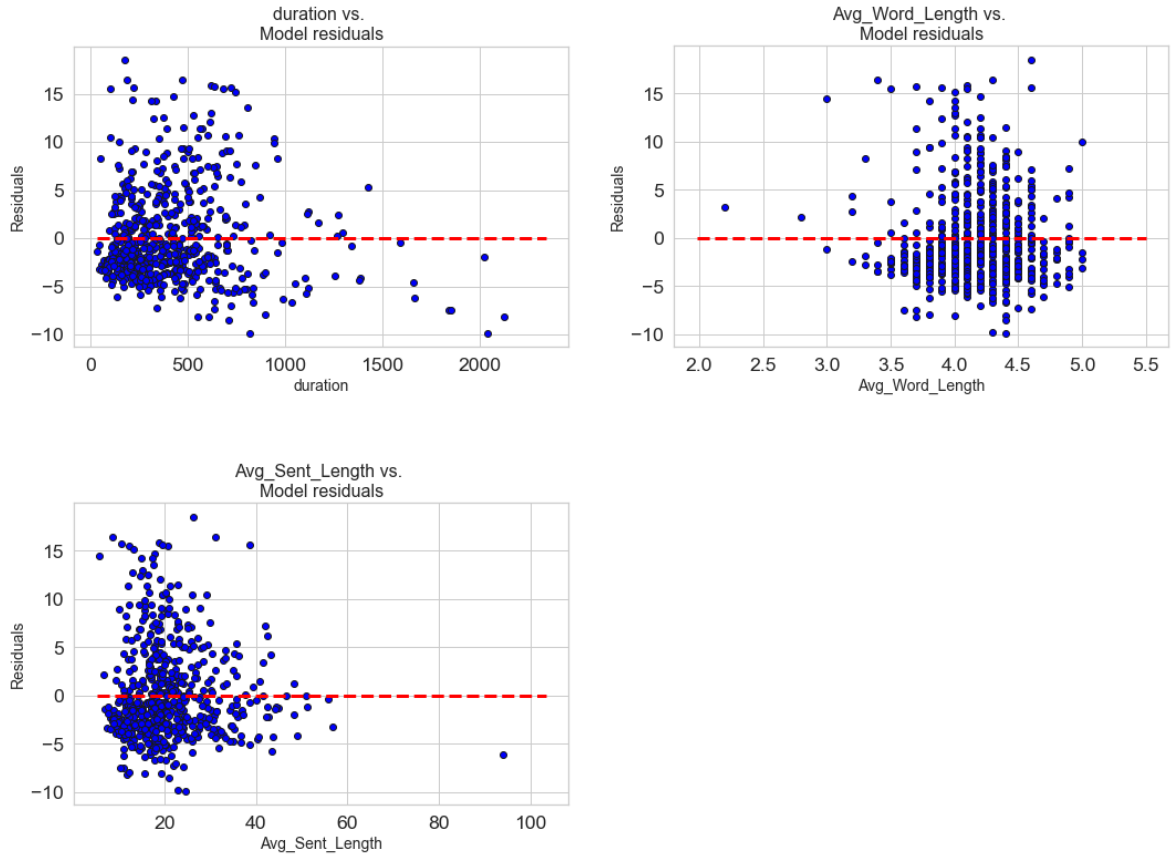
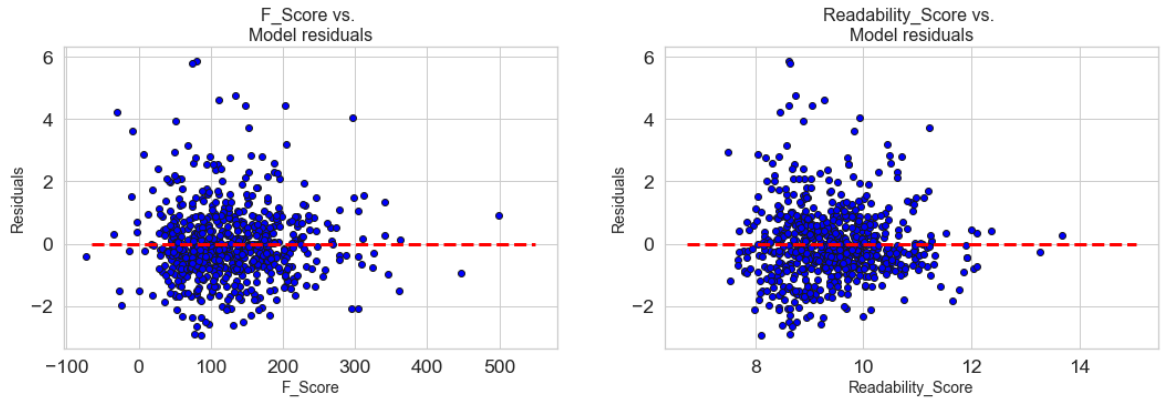


Figure 9.4 Residual Plots for the View Count Models

9.4.4.2 Engagement Rate

As can be judged from Figure 9.5, very similar patterns exist for the Engagement Rate as was seen for the regression based on view count, which is nice to see.



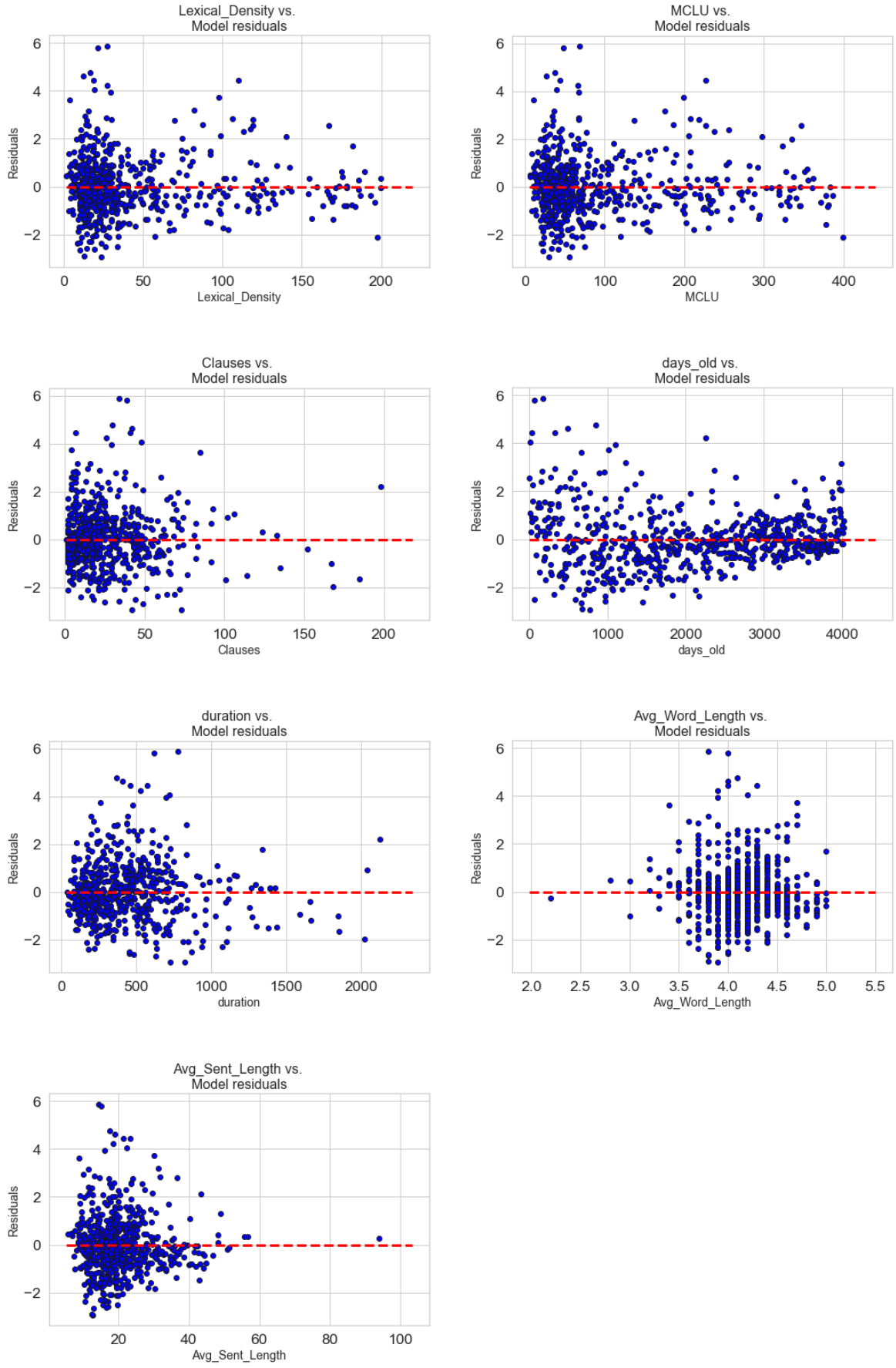
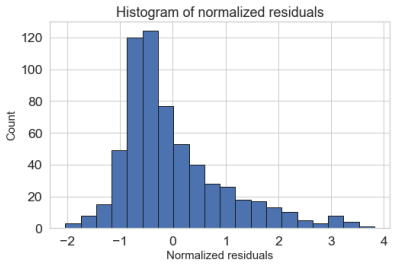


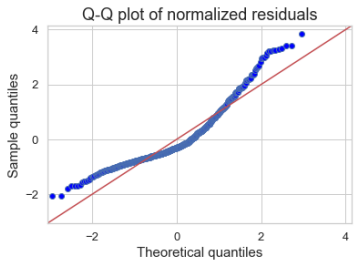
Figure 9.5 Residual Plots for the Engagement Rate Models

9.4.5 Normalized Residuals

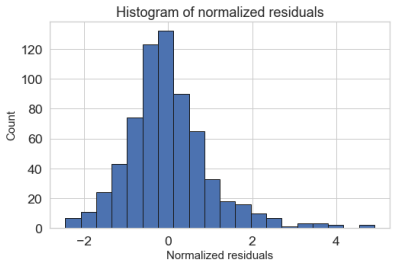
In order to check whether the residuals are normally distributed we produce a residuals histogram and Q-Q plot per each model. The residuals for the View Count models are slightly less normally distributed than those of the models for Engagement Rate, but not in any worrying sense, the same goes for the Q-Q plots.

View Count





Engagement Rate



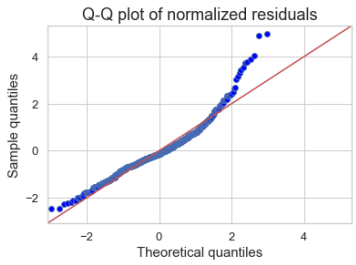


Figure 9.6 Assessment of Normalized Residuals

9.4.6 Homoscedasticity

To test for homoscedasticity, we deploy the Breusch-Pagan test. The null hypothesis states that homoscedasticity is present. The results are presented below in Table 9.5. Since either of our model variants return statistically insignificant results, we do not reject the null hypothesis; homoscedasticity is present.

Table 9.5 Breusch-Pagan Test Results

View Count	*Engagement Rate*
Test Statistics: 16.445 $p = 0.10573$	Test Statistics: 9.216 $p = 0.23598$

9.4.7 Cook's Distance for Outlier Detection

Cook's distance helps us identifying potential outliers that may impact the results of our model. The View Count models include one or two substantial outliers; however, since we already excluded a large portion of outliers, further data loss would not be worth the gain in explanatory power. The outliers in the Engagement Rate models do not demand further attention since they are merely minor.

View Count	*Engagement Rate*

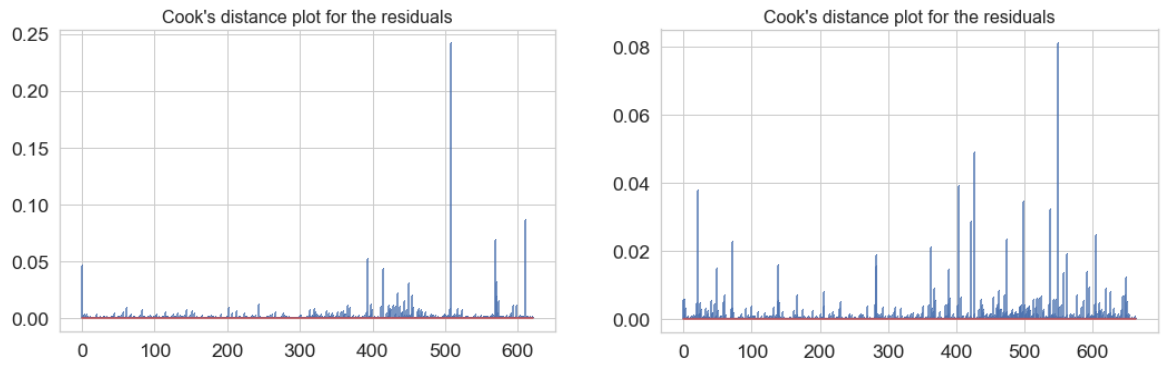


Figure 9.7 Cook's Distance Plot