Hackathon Idea: Youtube's automated speech recognition, closed captioning, and sentiment analysis from non-native English speakers

Duubar Villalobos Jimenez[a], Team member 1[b], Team member 2[b,*]

[a] *CUNY School of Professional Studies, New York, NY*
[b] *CUNY Baruch College, New York, NY*

Abstract

Would you think that YouTube's speech recognition technology is accurate for various accents? Would you believe that YouTube's automatic captioning change the expressed sentiment of the spoken words given by a non-native English-speaking person? The idea behind this research, strives to answer these questions. For this, we will interpret the definition of sentiment analysis as follows: "The process of computationally identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral." To answer these questions, we will seek two different hypotheses for each item. We will explore diverse YouTube videos with non-native English participants. Once we analyze the results, we can then draw our conclusions towards the end.

This is also part of my masters' capstone research project in Data Science in conjunction with Dipika Shrestha, masters' in Public and International affairs.

Keywords: YouTube, Automation, Speech recognition, Closed Captioning, Non-native, English, Sentiment Analysis, Technology, Communication.

---

*Corresponding Author

Email addresses: `Duubar.VillalobosJimenez@spsmail.cuny.edu` (Duubar Villalobos Jimenez), `cuny@email.com` (Team member 1), `cuny@email.com` (Team member 2)

1. Literature Review/Research Conducted

1.1. English becomes the global lingua franca

Globally, English has become a lingua franca or language of communication, and the number of users who are not native speakers has exceeded the numbers of native speakers.

English is spoken at a useful level by some 1.75 billion people worldwide. There are close to 385 million native speakers in countries like the U.S. and Australia, about a billion fluent speakers in formerly colonized nations such as India and Nigeria, and millions of people around the world who've studied it as a second language. An estimated 565 million people use it on the internet everyday.

Table 1: English speakers around the world

Speakers	Description
385 Million	Native English Speakers
1.365 Billion	Non-native English Speakers

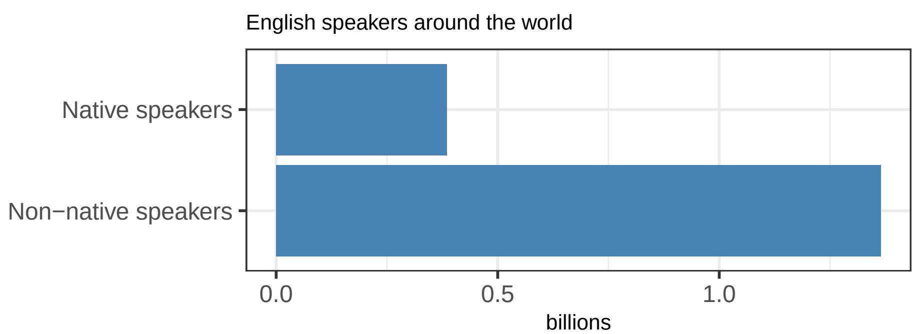

Figure 1: English speakers around the world.

Something interesting to note is that some studies express concerns. Such concerns indicate that the numerous non-native speakers who use English to communicate with other non-native speakers every day are affecting the English language.

1.1.1. *Languages with the most speakers*

Figure 2 shows how the top 4 languages are divided among speakers.

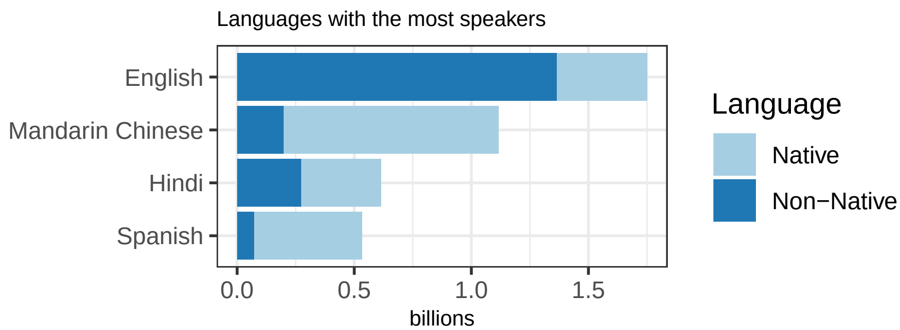

Figure 2: Languages with the most speakers.

1.1.2. *Top 10 most spoken languages*

The following, is a list of the 10 most spoken languages in the world, totaling 5.526 billion –representing around 72 %, of the 7.7 billion current world's population.

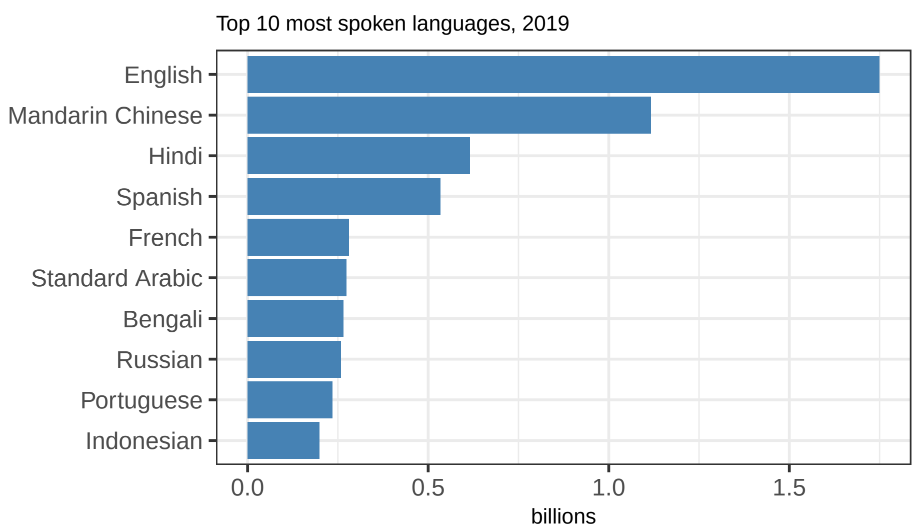

Figure 3: Top 10 most spoken languages, 2019.

1.1.3. *Becoming English proficient*

In a study performed by The University of California Linguistic Minority Research Institute, concluded that English academic proficiency takes longer to develop than oral English proficiency. The range for academic English proficiency development takes between four to seven years. The previous study seems to match an independent Dissertation submitted to the Faculty of the Graduate School of the University of Maryland. The study concluded that the median number of years that takes to reclassify non-native Female students as English proficient takes 3.7 years, compared to non-native Male students, who require a median of 4 years.

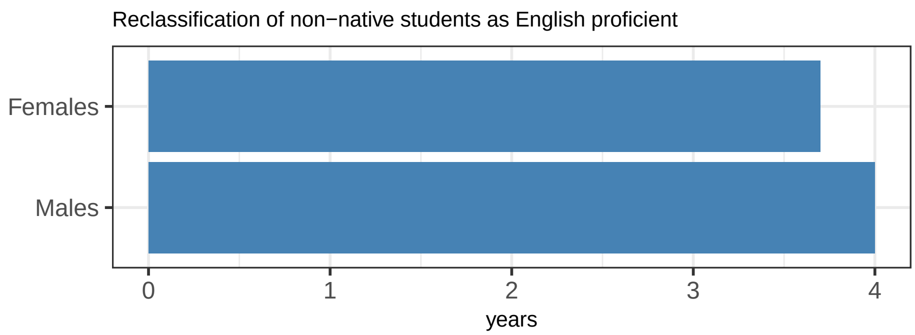

Figure 4: Median number of years to reclassify a non-native student as English proficient.

1.1.4. Constituency of words in the communication process

Alexander Arguelles provided the following constituency of words.

Table 2: Constituency of words.

Words	Constituency
250	Essential core of a language, those without which you cannot construct any sentence.
750	Used every single day by every person who speaks the language.
2500	Enable you to express everything you possibly want to say, albeit often by awkward circumlocutions.
5000	Active vocabulary of native speakers without higher education.
10000	Active vocabulary of native speakers with higher education.
20000	Needed to recognize passively to read, understand, and enjoy a work of literature such as a novel by a notable author.

1.1.5. English as the global language of business

The following table, shows how English is now the global language of business.

Table 3: Progressing from beginner level to advanced.

Words	Gauging Fluency	Description
250 to 1500	Beginner	Able to cope with basic situations.

Words	Gauging Fluency	Description
3000 to 5000	Intermediate	Able to understand verbal and written communications and express themselves.
5000 to 10000	Advanced	Able to communicate comfortably with technical terms and nuanced discussion.
10000 +	Native Speaker	Able to speak fluently idiomatically and have all means at their disposal to communicate effectively.

From table 2 and table 3, we can easily recognize some similarities.

From the above studies, we learned how English had become the preferred global language of communication. That is, "It's easier to speak broken English than a broken Mandarin." Also, we have learned as to how many years of the rigorous learning experience are needed to be considered a fluent English native speaker. We were able to learn and quantify the number of words needed to be be considered an English native speaker. Our focus now will center on technology, in particular, YouTube.

1.2. YouTube and Closed Captioning

Let's begin by referring to some valuable information about this popular platform. For this, we will make use of a small timeline to provide a series of important historical events.

Table 4: YouTube and Closed Captioning timeline

Date	Historic event
April 23, 2005	An 18-second clip about how cool elephants are was shot at the San Diego Zoo and uploaded to a then-private video sharing site called YouTube.
September 19, 2006	Google announced that they have just introduced a small but significant new feature that many of us have long-awaited. Playback with captions and subtitles on Google Video!
November 19, 2009	By then, Google had already acquired YouTube. Google announced the new automated captioning service on YouTube.

1.3. YouTube's key performance indicators

In this section, we will focus on some key indicators that make our research appealing, especially for which YouTube has become a Global phenomenon accessed by many.

Table 5: YouTube by the numbers

Date	News
March 20, 2013	1- Youtube hits one billion unique monthly visitors. 2- Nearly one out of every two people on the Internet visits YouTube.
May 01, 2013	Youtube connects 15 percent of the planet population to the videos they love.
October 12, 2015	More than 80 percent of YouTube's billions of views come from fans in countries outside the U.S.
February 16, 2017	1- The number of videos with automatic captions now exceeds a staggering 1 billion. 2- People watch a video with automatic captions more than 15 million times per day. 3- A 50 percent leap in accuracy for automatic captions in English has been achieved.
June 22, 2017	1.5 billion logged-in viewers visit YouTube every single month.
February 1, 2018	YouTube Go is available in over 130 countries around the globe.
June 21, 2018	1- More than 1.9 billion logged-in users who come to YouTube every month. 2- YouTube localized versions stretching across 90 countries and 80 languages.

1.3.1. YouTubes monthly visitors

In the figure 5, we can appreciate the number of montly unique visitors to YouTube.

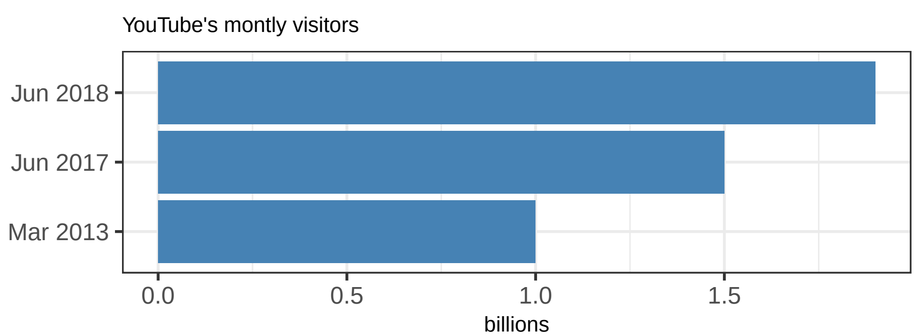

Figure 5: YouTube's montly visitors.

The above facts are essential to our study since they exemplify the magnitude of YouTube viewers' growth among non-English speakers.

1.3.2. Google's motto

"Our motto is to organize the world's information and make it accessible and useful." – "This is an innovation that takes advantage of our speech recognition technology to turn the spoken word into text captions."

I believe worth mentioning that over 60 accessibility leaders from the National Association of the Deaf, Gallaudet University, the American Association of People with Disabilities (AAPD), and other organizations, also joined Google to be the first to learn about these new features.

The automated captioning service was designed to help people who are deaf or hearing-impaired. From my perspective, this is an essential piece of information. And from here, moving forward, our analysis will determine if the current automated captioning is affected by different accents, and if the sentiment of the captured words from non-native English-speakers gets affected. We hope that our results and conclusions will become a determinant factor in order to help those deaf, hard hearing, or non-English speaking people who rely on the accurate captioning of the words, and it's translation to other languages.

1.4. Deafness and hearing loss

To extrapolate the above information, here are some key facts related to deafness and hearing loss. The World Health Organization made the following statement: "Disabling hearing loss refers to hearing loss greater than 40 decibels (dB) in the better hearing ear in adults and a hearing loss greater than 30 dB in the better hearing ear in children. The majority of people with disabling hearing loss live in low and middle-income countries."

Table 6: Deafness and hearing loss by the numbers.

Value	Description
34 million	The number of children worldwide that have disabling hearing loss.
432 million	The number of adults worldwide that have disabling hearing loss.
466 million	The number of people worldwide that have disabling hearing loss.
900 million	The estimated number of people worldwide that will have disabling hearing loss by 2050. Or one in every ten people.
1.1 billion	The number of young people (aged between 12–35 years) at risk of hearing loss due to exposure to noise in recreational settings.
One-third	The number of people over 65 years of age that are affected by disabling hearing loss.
750 billion	Estimated annual global cost in –U.S. dollars, of unaddressed hearing loss.

In figure 6 we can apprecate the visualization for the above presented values.

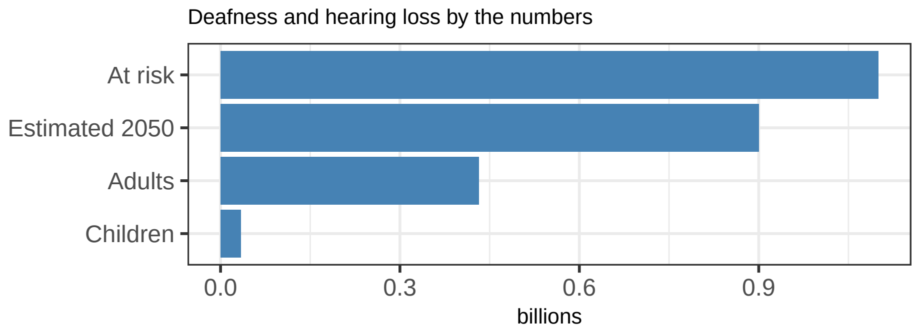

Figure 6: Deafness and hearing loss.

1.5. *Low and middle-income countries*

We have learned that the presence of hearing loss happens predominantly in low, and middle income countries; which by the way, are non-native English speakers in its majority. The above statement can be further confirmed by looking at the appended table in the next section, obtained from The World Bank.

1.5.1. *World Development Indicators*

World Development Indicators (WDI) is the primary World Bank collection of development indicators, compiled from officially recognized international sources. It presents the most current and accurate global development data available and includes national, regional, and worldwide estimates.

Table 7: Countries grouped by income. Last Updated: 10/02/2019

Low income	Lower middle income
Afghanistan	Angola
Benin	Bangladesh
Burkina Faso	Bhutan
Burundi	Bolivia
Central African Republic	Cabo Verde
Chad	Cambodia
Congo, Dem. Rep.	Cameroon
Eritrea	Comoros
Ethiopia	Congo, Rep.
Gambia, The	Cote d'Ivoire
Guinea	Djibouti
Guinea-Bissau	Egypt, Arab Rep.
Haiti	El Salvador
Korea, Dem. People's Rep.	Eswatini
Liberia	Ghana
Madagascar	Honduras

Low income	Lower middle income
Malawi	India
Mali	Indonesia
Mozambique	Kenya
Nepal	Kiribati
Niger	Kyrgyz Republic
Rwanda	Lao PDR
Sierra Leone	Lesotho
Somalia	Mauritania
South Sudan	Micronesia, Fed. Sts.
Syrian Arab Republic	Moldova
Tajikistan	Mongolia
Tanzania	Morocco
Togo	Myanmar
Uganda	Nicaragua
Yemen, Rep.	Nigeria
	Pakistan
	Papua New Guinea
	Philippines
	Sao Tome and Principe
	Senegal
	Solomon Islands
	Sudan
	Timor-Leste
	Tunisia
	Ukraine
	Uzbekistan
	Vanuatu
	Vietnam
	West Bank and Gaza
	Zambia
	Zimbabwe

And by considering the above information –once we link the need and the service, we find a lively "symbiotic" relationship between YouTube, people with hearing accessibility needs, and non-English speakers.

From the above-presented facts and timeline, we learned of Google's initial intent to create and promote the use of it's automated closed captioning service on YouTube. Primarily, directed to people who present some hearing disabilities, or those who are not proficient in English, by providing automatic translations as well. It is important to recall, that these services were also promoted for non-English speakers as well. Furthermore, the above findings provide an excellent idea for the number of people who are non-native English speakers. The number of people that have or will experience some hearing loss and the number of people currently relying on the automated YouTube captioning service around the world.

1.6. Spoken Language Recognition (SLR) technology

In the previous sections, we learned about how English became the global language of business. Also, we learned about English proficiency and what it takes to become a non-native English speaker. Moreover, we learned about YouTube's history and services. In this section, we will focus on the analytical part of one of the problems at hand. That is, How accurately does YouTube turn the spoken words into text captions for various English foreign accents?

Spoken Language Recognition (SLR) is the task of recognizing by computational means the language spoken in an utterance –a spoken word, statement, or vocal sound. Typically, SLR has been used as an auxiliary module in many applications, such as multilingual conversational systems, spoken language translation, and multilingual speech recognition for example. As we now know, Google has introduced this technology into YouTube videos to extract the spoken words and convert them into text.

1.6.1. YouTube and unstructured data

It is important to note that since the mid-2000S, business topics related to big data business analytics and unstructured data have received a lot of attention. Some examples are seeing when companies start analyzing data derived from social media, blogs, and email messages. In our case, we will be analyzing unstructured data from YouTube video sources. On those videos, non-native English speakers, share inspiring social stories to help newly arriving immigrants, navigate the nuances that they face on arrival to the United States. We just unlocked a valuable piece of information, because –as we now know, the main goal for SLR systems is to capture, categorize, store, and help to analyze unstructured data. In theory and practice, this process can be customized for each video to include language identification, audio entity extraction, and real-time monitoring. And by keeping that in mind, now we can process information hidden in the unstructured data related to immigration issues for example.

In a research study published by the Laboratory for Computer Science, Massachusetts Institute of Technology, back in the year 2000. The authors expressed, "Spoken language understanding involves the transformation of the speech signal into a meaning representation that can be used to interact with the specific application back-end. Two steps are the average number of processes needed to accomplish: 1- the conversion of the signal to a set of words (i.e., speech recognition). 2- the derivation of the meaning from the word hypotheses (i.e., language understanding)." Something interesting to note is that SLR systems need to take into consideration non-native speech in multilingual systems to achieve this goal.

Another essential piece of information is to make note that on February 16, 2017, YouTube reported a 50 percent leap in accuracy for automatic captions in English had been achieved. Another critical piece of information has just given light. That is, YouTube's press release did not express if the achievement of this milestone was for native English speakers only. Or if the results include non-native English speakers with foreign accents, as shown in the previous studies.

1.6.2. Characteristics of foreign-accented speech

A study performed by the Interactive Systems Laboratories, Karlsruhe University, Germany, in conjunction with the Carnegie-Mellon University, Pittsburgh, PA, pointed out that some of the characteristics of foreign-accented speech are: 1. Phoneme realization –stress patterns and durations play. 2. Articulation of phonemes in context. 3. Phonotactic constraints –different languages allow different sequences of phonemes. Also, another study pointed out that acoustic modeling for non-native speech must handle non-native models, bilingual models, model merging, and dictionary modification; thus, to significantly improve recognition of non-native speech.

Something interesting to note is that some researchers have significant concerns. Some of these concerns, relate to a lack of resources, to objectively assess SLR technology on other types of speech (e.g. speech produced by multiple speakers in different and changing environments), on separate sets of languages (e.g. European languages) or applications (e.g. indexing of the spoken language in spoken documents). Keep in mind that we will be working exclusively with non-native English spoken videos. We are assuming that YouTube trains their models in such ways. We are also assuming that YouTube improves the accuracy –of their SLR algorithms, with non-native English speakers with foreign accents as part of the training data-set.

From myr perspective, this topic is very vast and productive for discussion and research that is out from our initial scope. We want to point out that many approaches are taking place. Researchers are currently discussing diverse strategies. Worldwide discussions are now taking place, seeking to address non-native model training beyond the accent. Perplexity discussions –inability to deal with or understand something complicated or unaccountable, are also accounted. Other centers of discussions focus on the use of frequent trigrams –often used in natural language processing for performing statistical analysis of texts. Others focus on disfluencies –when the speaker is searching for the right word, expression, or is pronouncing a word that is difficult to articulate in spontaneous speech.

1.6.3. Noisy speech

Another aspect to consider is the noisy-speech segments. Noisy or overlapped speech may interfere with short fragments of clean speech. Different and variable types of noise may appear, some could be: street, music, cocktail party, laughs, clapping, etc. Most expression overlaps appear in hot spots of informal debates in late-night shows, magazines, or in our case, a podcast. For our study, we will feature clean-channel and quiet-background (studio) conditions to avoid getting into higher studying grounds of SLR technology.

1.7. Sentiment analysis

Now that we have a better understanding of Spoken Language Recognition technology. We have learned as to how it serves our purpose; the above explorations, help us to understand the way YouTube algorithms extract text from spoken words present on the video. Now, we will focus our literature review into the second part of our analysis, that is: Does automatic captioning change the expressed sentiment of the spoken words?

"Sentiment analysis is the computational study of people's opinions, sentiments, emotions, and attitudes. This fascinating problem is increasingly important in business and society. It offers numerous research challenges but promises insight useful to anyone interested in opinion analysis and social media analysis". The above, provides an introduction to the topic from a primarily natural-language-processing point of view. The main idea is to help understand the underlying structure of the problem and the language constructs that are commonly used to express opinions and sentiments. It is also worth mentioning, that core areas of sentiment analysis, includes many emerging themes, such as debate analysis, intention mining, and fake-opinion detection. The main focus is to employ computational methods to analyze and summarize opinions. It is said that this area of study offers valuable resources for researchers and practitioners in natural language processing, computer science, management sciences, and the social sciences.

1.7.1. Sentiment vs Sentiment analysis

For this research, we consider it very important to clarify a few things. First, what is sentiment? And who's sentiment are we referring? To answer those preliminaries, we are going to make use of the following definition "**Sentiment:** 1. A view or opinion that is held or expressed. 2. General feeling or opinion. 3. A feeling or emotion." So, putting it together, we could say that sentiment is… "a general view, opinion, feeling or emotion that is expressed." Now, let's analyze the following definition, "**Sentiment Analysis:** Is the process of *computationally identifying and categorizing* opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral." From the above, We can conclude that sentiment and sentiment analysis are two different things even though similar descriptions are given. The sentiment is an *expression* of a general view, opinion, feeling, or emotion. Sentiment analysis is a *process* in which computers use mathematical algorithms to determine if the textual attitude towards a particular topic is positive, negative, or neutral. This is a very important piece of information for someone who's not familiar with the term sentiment analysis. Neutral usually means no opinion.

To extrapolate the above definition and with hopes of clearing any misunderstanding, investigators from The Institute of Automation, Chinese Academy of Sciences expressed, "This fascinating problem is increasingly important in business and society." Their comments, also make reference to some concerns that many of us might have, "Although we have known sentiment analysis as a task of mining opinions expressed in text and analyzing the entailed sentiments and emotions, so far the task is still vaguely defined in the research literature because it involves many overlapping concepts and sub-tasks. Because this is an important area of scientific research, the field needs to clear this vagueness and define various directions and aspects in detail, especially for students, scholars, and developers new to the field."

1.7.2. Different levels of analysis

Currently, these are the recognized sentiment analysis levels:

Table 8: Sentiment analysis levels.

Level of Analysis	Description
Document-level	Classify whether a whole opinion document expresses a positive or negative sentiment. Commonly known as document-level sentiment classification.
Sentence-level	Classify whether each sentence expresses a positive, negative, or neutral opinion. Widely known as subjectivity classification.
Entity and aspect-level	Performs finer-grained analysis. The aspect level was earlier called the feature level.

It is interesting to note the presence of many great examples in the book. Something worth mentioning, is that those examples are grammatically correct. However, in our case, there's no direct presence of well-made sentences, thus making our analysis a little bit more challenging.

1.7.2.1. Sentiment lexicon and its issues. Not surprisingly, the most critical indicators of sentiments are sentiment words, also called opinion words. These are words that are commonly used to express positive or negative views. For example, kind, beautiful, and marvelous are positive sentiment words, and evil, weak, and terrible are negative sentiment words. Apart from individual words, there are also phrases and idioms, e.g., "cost someone an arm and a leg." Sentiment words and phrases are instrumental to sentiment analysis for obvious reasons. A list of such words and phrases is called a sentiment lexicon –or opinion lexicon.

1.7.2.2. Natural Language Processing (NLP). Sentiment analysis is an NLP problem. "It touches every aspect of NLP, e.g., coreference resolution, negation handling, and word sense disambiguation, which add more difficulties since these are not solved problems in NLP. However, it is also useful to realize that sentiment analysis is a highly restricted NLP problem because the system does not need to understand the semantics of each sentence or document fully but only needs to understand some aspects of it, i.e., positive or negative sentiments and their target entities or topics. In this sense, sentiment analysis offers a great platform for NLP researchers to make tangible signs of progress on all fronts of NLP with the potential of making a huge practical impact."

1.7.2.3. Supervised learning methods and machine learning. Another important aspect of our research, is that supervised learning methods provide no linguistic interpretations, and no knowledge is generated for linguists, or industry developers, to gain insights into the problem. When errors occur in an application, it is hard to know what is wrong and how to fix it. "Fortunately, there are a lot of comprehensive list of linguistic constructs and perspectives that are instrumental for sentiment analysis, which make up for the deficiency of black-box approaches using pure machine learning. Moreover, it also lists and elaborates on many specific linguistic phenomena that are critical for effective classification of sentiments such as negation, modality, and comparison. We believe that the work presented by Liu, will enable us to gain a comprehensive understanding of the computation methods, deep linguistic insights of the sentiment analysis problem, and its possible solutions" From my perspective, this is very important for us. We strive to perform analysis that we believe has not been done before, and it might ignite some other research in the future.

The following is an essential piece of information "Although many sentiment analysis methods concentrate on machine learning –as in other NLP tasks, sentiment analysis is much more than just a classification or regression problem, because the natural language constructs used to express opinions, sentiments, and emotions are highly sophisticated, including sentiment shift, implicated expression, sarcasm, and so on." We will work exclusively on sentiment analysis, and we must keep these critical observations in mind.

1.7.3. Importance of sentiment analysis

As I start to wrap up, we have shown that sentiment analysis –also known as opinion mining, refers to the problem of identifying the dominant sentiment in a given piece of text. The sentiment is usually modeled as a categorical variable with three values: positive, negative, and neutral. Every day, with the ever-increasing need to process information, the evaluation of the sentiment present in pieces of text –in our case, video, help to identify better and analyze the minds of the people —usually to make better policy decisions, be it in business or government.

1.8. English around the world

Figure 7 shows how English has spread around the world.

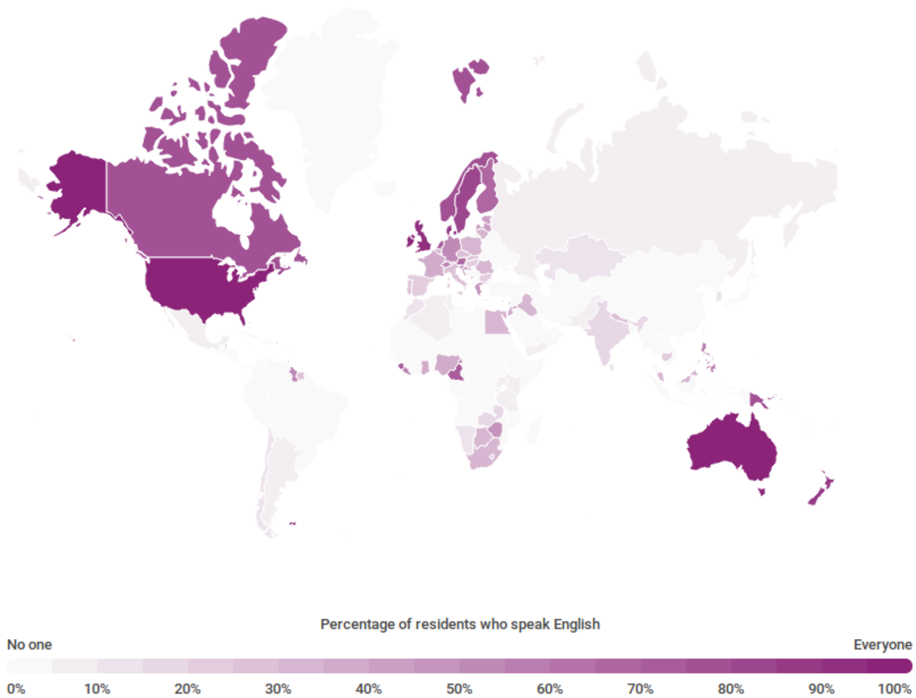

Figure 7: English around the world.

2. Results

In this section, we could present a table with the following results.

Table 9: Initial study.

Video ID	Gender	Accent	Country	Years in the U.S.

Table 10: Final results study.

Video ID	Number of Utterance	Words in Auto captioning	Words in Correct Transcript	Number of Error Words	Error Percentage

3. Discussion

We could present for discussion.

4. Conclusion

The center of our research is not on the prevention or treatment of hearing loss. Our study focuses on the promotion of social inclusion for people with disabilities. It also includes people with hearing loss and deafness. Also, it includes people who are not fluent English speakers, since they rely on the automated captioning and the automated translation service provided by YouTube.

We believe that this study will help gain the confidence to those who can't listen or speak fluent English. With this study, we aim to prove scientifically, how accurate the sentiment of a non-native English speaker given message is kept by the automated captioning technology provided by YouTube.

5. Appendix

References