Performing Key Data and Semantic Analysis on Bimodal Datasets

Victor Corja, Austin Crow, Nannan Liu

1. Problem Statement

The main challenge faced by data analysts after any significant event in which public sentiment is a key factor of analysis is the volume of information available. An occurrence affecting a large group of people, such as some form of a natural disaster, or an election, results in the creation of massive amounts of data without any standard format or medium, and this makes the job of the data analyst that much harder. We would like to identify the most effective models currently developed for the semantic evaluation of text and speech data and provide a model of our own which would permit the input of any text and speech datasets, separately or combined, and return a comprehensive overview of the key data points and semantic identifiers contained within. This model would allow future researchers to bypass the issue of determining which medium to focus on, as the metadata analysis within would combine the efficacy of existing text and audio data parsing algorithms.

Our initial evaluation process is based on a literature review of existing proposals and ratings of models by prior researchers, which will guide our decisions as to the capabilities built into our final deliverable. Having determined the general composition of our model, we would then begin a training process with a number of text and speech datasets of publicly generated data revolving around key events within the past decade, through which we would determine the efficacy and any potential requisite alterations required of our model. Finally, we will perform a comprehensive analysis of the capabilities and limitations of our model, as well as future work required to increase its effectiveness in a greater range of situations.

2. Objectives

There are two primary objectives for this research process. The first is to identify the leading models used for data parsing, metadata analysis, and semantic discrimination of text and audio data. The second is to propose a model which would capitalize on the best existing models to generate a report identifying the key topics and semantic patterns from any dataset composed of text and/or audio resources. The model's operative process would be as follows:

1. Combine speech and text resources into one resource
2. Determine topics and keywords from the resource
3. Identify the semantics and mood from this resource
4. Generate a report with a detailed analysis of the obtained information

The results obtained from this assessment can provide significant benefits to professionals in the news, emergency/incident response, and political analysis spheres. By reviewing the key themes and semantic undertones of data generated by social media users regarding a specific viral topic or political event, analysts would be able to better understand the public reaction and potentially better plan or prepare for the next such occurrence. Similarly, investigators analyzing data in the aftermath of a natural disaster may better be able to understand the impact by investigating compiled social media posts and calls made to the emergency services.

3. Literature Review

In order to separate our analysis of the existing literature in our field of research, we split the review into three sections, each focusing on a particular aspect of our paper. The first section is focused on publications which discuss the analysis of audio resources, including through conversion to text and direct analysis. The second section identifies papers in which methodologies are proposed for the identification of key data in text resources, such as topic mining and keyword analysis. The third section is focused on research into the semantic analysis of data, including the determination of key themes and moods within text and audio data.

3.1 Metadata Identification within Audio Inputs

Text data is the main subject of analysis within the majority of research, largely due to its availability and relative ease of analysis. There are hundreds of models available which can perform semantic analysis, deep mining, geotagging, and other evaluations based on a set of textual inputs. However, considering the enormous amount of audio data generated via social networks, telephone calls, video sharing platforms, and other outlets, as well as increases in processing power and developments of new sensory models, affective analysis can be performed on multimodal data such as audio files, generating volumes of new metadata (Shoumy et al. 2020). Audio mining, a technique to search through an audio resource and identify high value information, can be used to analyze audio, as well as composite audio, video, and/or text data for semantic cues, keywords and subjects, as well as identifying certain flags within the data such as whether it contains lies or indicators of hate speech (Shoumy et al. 2020). Our research looked into both the possibility of directly analyzing audio data, and that of combining speech and text resources. While direct analysis would potentially be more effective for audio resources, there is the question of complexity, which would be greatly reduced if we could perform the analysis on a set of data in a consistent medium.

If we converted audio data into text information, we could then perform word segmentation, information classification, and extraction. Combining various classification models and neural network models to analyze a unimodal data set could lead to a model with greatly increased efficiency for sentiment analysis. There are many methods of performing speech recognition on audio data, a few of which are listed and ranked by Toshniwal et al. (2018). The authors suggest that combining the traditionally separate automatic speech recognition (ASR), learned acoustic model, pronunciation model, and language model (LM) into the same single network is the best and most effective way of working with speech data. The focus of the paper is on the differences between the language models, which can be categorized into shallow, deep, and cold fusion, and have different integration timings and training times. Per the authors' analysis of the models based on tests performed on two datasets, it was determined that shallow fusion is generally the best approach until the "second pass rescoring", in which cold fusion takes the lead.

The alternative approach to analyzing audio inputs, and one that is preferable when it is important to preserve speech pattern metadata (accents, inflections, etc.). Tschöpel and Schneider (2010) discuss the Fraunhofer IAIS AudioMining system, which performs vocabulary-independent spoken-term detection. This technology determines keywords from audio inputs, storing the spliced audio as opposed to a text index. In order to enhance the functionality of this model, they introduced a neighborhood algorithm, which separated the speech into different topics, similar to the more well-known text topic mining techniques. Lu and Hanjalic (2006) discussed the use of audio-elements within audio data analytics, a mid-tier structure between low- and high-level features in the audio input such as applause, whistling, or the sound of an impact in a soccer match. The proposed approach used is taking an audio document, discovering the audio elements via spectral clustering, determining the Elements Similarity Measure and then performing term and element weighting. The results of their analysis showed an increase in the performance of their auditory scene segmentation model as a result of using audio-elements.

3.2 Metadata Identification within Text Inputs

Information extraction refers to extracting entity and relationship attributes from text and speech metadata, which is convenient for retrieval, query, and analysis. As an example, information extraction related to a natural disaster would retrieve the time, location, and overview of the situation from a basic data set into a structured data set to simplify data management and data analysis. In order to perform a comprehensive analysis of key data identifiers within the textual components of our datasets we looked at the methods of extracting information from text resources which we could implement within our model.

One commonly used approach to text analysis is topic mining, which involves the process of grouping input data into clusters by using a similarity index. “Topic mining as a scientific literature can accurately capture the contextual structure of a topic, track research hotspots within a field…” (Zhang et al 2020). By grouping key features from the data, the clusters then can be quantified in the number of relationships there are between topics and features, thus giving a strong visual analysis of what the data is about, all done with limited loss of the textual implications in the data.

Lee and Kim (2008) describe the use of term frequency (TF) in metadata analysis, an identification of a word or word pattern that appears most frequently in the article, while ignoring common stop words - terms that do not add to the value of the article. The authors implemented an importance adjustment coefficient to measure whether a word is contextually relevant, using the Inverse Document Frequency (IDF) as the weight of the commonality of a term. The product of the TF and IDF is equivalent to the importance of the word within the article, and this method has the advantage of being simple and fast, and the result is more in line with the actual situation.

Rose et al. (2010) … wrote about the Rapid Automatic Keyword Extraction (RAKE) algorithm, which extracts key phrases from text inputs. The RAKE algorithm first uses punctuation to divide a document into several phrases, then uses stop words as separators to divide the phrase into sub phrases, which are candidates for the final extracted key phrases. Each phrase can be further divided into several words by spaces, and each word can be assigned a score to each word by calculating the quotient of the degree of significance and frequency of appearance.

Zhang et al. (2020) propose the TextRank algorithm is a graph-based ranking algorithm for text. The basic idea comes from Google's PageRank algorithm. By dividing the text into several constituent units (words, sentences) and building a graph model, the sentences are sorted, and keyword extraction can be realized only by using the information of a single document. It is widely used because of its simplicity and effectiveness.

3.3 Semantic Analysis of Text and Audio Inputs

Semantic analysis is the process in which a computer understands the sequence and meaning of words in the same way a human would, including a contextual understanding of colloquialism and homographs. In the past decade, deep semantic analysis of datasets has become possible with machine learning enabling models to classify metadata based on contextual indicators found within. Semantic classification technology plays an important role in intelligent information processing services, identifying themes within the data and increasing the metadata which can be extracted from collected data. As with the prior research, we analyzed developments for semantic analysis within both text and audio inputs.

Semantic analysis of audio inputs is rather complicated to do as machines do not have emotion therefore cannot gauge emotion inherently. Audio-Speech Recognition (ASR) has helped with this issue by being able to break down audio into data that the computer can process. This is commonly done using Mel-frequency cepstral coefficients (MFCC) values. In the article by Yoon et al (2018) they propose a model that is more accurate than the two previous model’s audio recurrent encoder (ARE) and text recurrent encoder (TRE). Their model is called the Multimodal Dual Recurrent Encoder (MDRE). In their model

they work on emotion prediction by using the actual sounds as well as the context. As well as combining high level text transcription and with low level audio signals. Feature extraction was performed using the aforementioned MFCC values and the model's performance was based on the average weighted precision in a 5 cross-fold validation. They concluded that "MDRE model compensates for the weakness of the previous two models and benefits from their strengths" (Yoon et al, 2018). This project is incredibly useful as it is accurately determining the mood of the data between a set scale thus showing that computers can read emotion.

There has been a significant amount of research performed on semantic analysis of text-based data, for a number of different purposes. Foltz (1996) writes about the use of latent semantic analysis, a statistical model that can predict the similarity between two inputs, for the purpose of determining the original sources of a researcher's summary. Latent semantic analysis functions by creating matrices of the occurrences of words in relation to other words and analyzes the percentage match between the occurrences across inputs. Another use for latent semantic analysis determined by Steinberger and Ježek (2004) was to identify important semantically important sentences within a text input, which they created by applying Singular Value Decomposition (SVD), creating a matrix of terms and sentences, and identifying the relative importance by repetition of appearance of a word pattern within the matrix.

4 Proposed Methodology

Prior to the beginning of our investigation, our proposed methodology consists of three key steps.

The methodology for this assessment will begin identifying 20 existing tools. These tools will then be compared against one another in their respective categories. The models that perform the best will then move forward in our development process, and will be selected from:

- 5 speech recognition models
- 5 audio mining models
- 5 key data mining models for text data
- 5 semantic analysis models for text and audio data

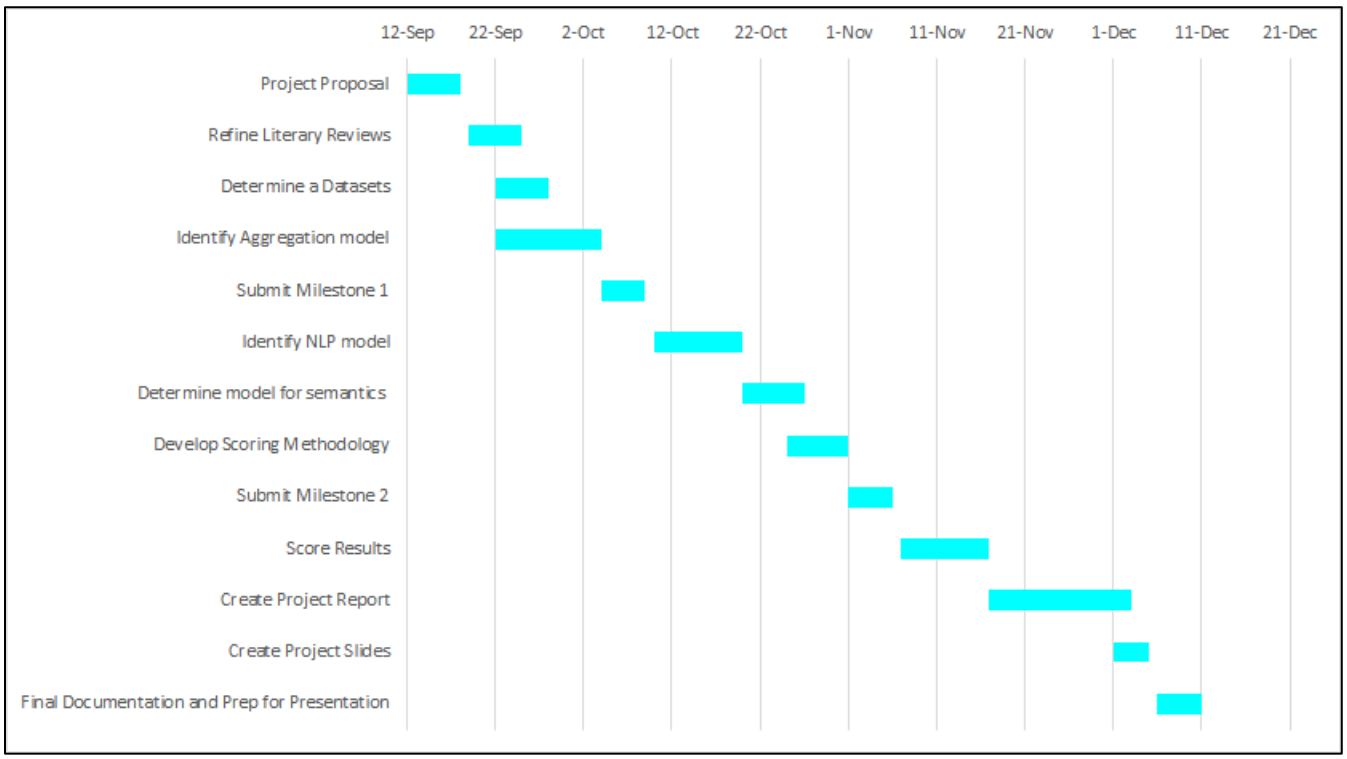
The second step would be to perform tests on the chosen models, likely using scripts built in Python with a set of datasets including both text and audio data, which we would break up into 20% to be used as a training set, and 80% as an analysis set. Through this analysis, we would decide on the algorithms and methodologies we would use in our final proposed model. The third step would be to perform a complete analysis of a new set of datasets focused on a particular event, which would allow us to make an evaluation of our approach and determine its areas of strength and weakness.

Once the best performing models are decided we will implement these models into a pipeline in which they all flow and work together. The model will be trained on the dataset via python, breaking up our set into a training set and test set, 20% and 80% respectively.

Our dataset(s) will require a fair amount of pre-processing as they are likely to be formatted in dramatically different ways. The preprocessing will follow best practices depending on the models chosen. All pre-processing will be done keeping in mind not to change any of the actual data to keep sets valid and unbiased.

Through testing the accuracy of these models, we can determine what algorithms will likely be the best, and which models to base our deliverable on. After this testing is complete, we will take this model and perform an analysis of a new dataset that focuses on a particular well-known event, thus, allowing a proper evaluation of the approach. This will enable us to identify the comparative strengths and weaknesses of our approach, as well as the output differences versus the performance of the existing models.

5 Project Development Timeline (Gantt)



References

Ali, M. M, and L Rajamani. "Deceptive Phishing Detection System: From Audio and Text Messages in Instant Messengers Using Data Mining Approach." *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*. IEEE, 2012. 458–465. Web.

Foltz, Peter W. "Latent Semantic Analysis for Text-Based Research." *Behavior research methods, instruments, & computers 28.2* (1996): 197–202. Web.

Lu, Lie, and Alan Hanjalic. "Towards Optimal Audio 'Keywords' Detection for Audio Content Analysis and Discovery." *Proceedings of the 14th ACM International Conference on Multimedia*. ACM, 2006. 825–834. Web.

Rose, Stuart et al. "Automatic Keyword Extraction from Individual Documents." *Text Mining*. Chichester, UK: John Wiley & Sons, Ltd, 2010. 1–20. Web.

Shoumy, Nusrat J et al. "Multimodal Big Data Affective Analytics: A Comprehensive Survey Using Text, Audio, Visual and Physiological Signals." *Journal of network and computer* applications 149 (2020): 102447–. Web.

Steinberger, J, and K Jezek. "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation." (2004).

Sungjick Lee, and Han-Joon Kim. "News Keyword Extraction for Topic Tracking." *2008 Fourth International Conference on Networked Computing and Advanced Information Management*. Vol. 2. IEEE, 2008. 554–559. Web.

Toshniwal, Shubham et al. "A Comparison of Techniques for Language Model Integration in Encoder-Decoder Speech Recognition." (2018): n. pag. Print.

Tschöpel, S, and D Schneider. *A Lightweight Keyword and Tag-Cloud Retrieval Algorithm for Automatic Speech Recognition Transcripts: Presentation Held at the 11th Annual International Conference on Spoken Language Processing, (INTERSPEECH), Makuhari, Japan, 26.-30.09.2010*. N.p., 2010. Print.

Yoon, Seunghyun, Seokhyun Byun, and Kyomin Jung. “Multimodal Speech Emotion Recognition Using Audio and Text.” (2018): n. pag. Print.

Zhang, Tingting et al. “Multi-Dimension Topic Mining Based on Hierarchical Semantic Graph Model.” *IEEE access 8* (2020): 64820–64835. Web.

Zhang, Mingxi et al. “An Empirical Study of TextRank for Keyword Extraction.” *IEEE access 8* (2020): 178849–178858. Web.