Bimodal Data Mining: Integration of Key Data and Semantic Analysis for Text/Audio Datasets

Victor Corja, Austin Crow, Nannan Liu

I. 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.

II. 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.

III. Literature Review

In our literature review, we identified existing research into various methods of key data extraction and semantic analysis, for both audio and text resources. Our focus was on determining models we could integrate into our own research, and for this purpose we selected studies which utilized open-source models, and had significant documentation. 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.

A. The first section is focused on publications which discuss models enabling the analysis of audio resources through the conversion of the audio data into text. This will enable us to later perform keyword extraction and semantic analysis on the total combined data. Our initial research looked into the possibility of directly analyzing audio data, as well as converting the audio resources into text. While direct analysis would potentially be more effective for audio resources, considering the complexity of existing approaches, as well as the significant difference in both the requirements and the outputs of those approaches, we decided to focus only on integrating models dealing with speech recognition for audio-to-text conversion.

B. 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 extraction. While data analysis is usually performed on structured data, in order to retrieve all possible relevant information we will need to perform our analysis on the unstructured text data contained within our datasets. This will require data pre-processing in order to account for the enormous amount of information, followed by keyword analysis and topic mining, both common approaches further identified within section III.b.

C. 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. Data is composed of words, sentences, and paragraphs, so semantic analysis can also be divided into lexical-, sentence-, and paragraph-level semantic analysis to provide insight into the emotional and contextual composition of text resources. These analyses determine word sense disambiguation, the links between text entities such as location, time, and reason, and overall themes within the text, respectively.

III.a 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 [1]. 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 [1].

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 [2]. 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.

Hidden Markov models (HMM) are one of the oldest and most prevalent models for speech recognition, used for sequence analysis. Prior to the use of the model, feature extraction is required, primarily in the form of Mel Frequency Cepstral Coefficients (MFCC), since Markov chains require discrete states. After using MFCC, the extracted features are converted into discrete variables, and can be analyzed using HMM, which uses a generative probabilistic model to determine the next character based on the relationships between two sets of variables. In order to identify the next most likely character, the model requires an input with specific information about the language, which can be used for training purposes.

Listen, Attend, Spell (LAS) is an end-to-end model for automatic speech recognition, differing from HMM in that it does not make assumptions about the output sequence. LAS works by transcribing the audio sequence signal to a word sequence one character at a time. The operation is performed in two sequences, the "Listen" and the "Attend and Spell" operations. The first operation transforms the original signal into a high-level representation," while the second takes the high-level representation and produces the probability distribution over character sequences [4].

The third model we looked at for speech recognition used Recurrent Neural Networks (RNN), a variation on standard neural networks focusing on differentiating phonemes. RNN does not require prior training in the language being analyzed, making it far more adaptable than models such as HMM. RNN focuses on networks with multiple feedback connections, creating nodes that contain deep-seated memory which can be queried [5]. By using backpropagation, each network can iterate through the provided data and create weights for likely values which can be shared across networks.

III.b 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…" [6]. 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. Another approach to text analysis, described by Lee and Kim [7], uses 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.

The first model we considered for our research was probabilistic data mining, described by Wang et al. [8]. The model involves a combination of topic mining and term frequency analysis and begins by creating a matrix of unique words within a dataset, then removing the highest and lowest frequency terms based on their TF-IDF product to account for both extremely common words like “and” and highly uncommon words. Once the matrix has been normalized to account for term frequency outliers, the authors then used the software R to randomly choose a distribution over topics and determine the topic proportions for certain topics within each document in the dataset. Each topic was randomly assigned to every n^{th} word in each document, resulting in “the high-probability terms that define a topic in the corpus”, and once the model had determined 20 topics for each document, the process was repeated a total of 10 times, diminishing the impact of determining topics from a lower-frequency word.

Zhang et al. [9] proposed the TextRank algorithm, which we selected as our second possible text analysis model. The TextRank model is a graph-based ranking algorithm for determining term relationships within text-based datasets, based on Google's PageRank algorithm. The modelling begins by displaying the significant words within the dataset as nodes, and creates edges between the nodes based on the degree of correlation within a close proximity. A TextRank score is then computed, based on the number of correlations and the significance decay by order of correlation, and the list of nodes with the highest scores is returned as the words of highest importance.

The third model we selected for analysis was the Rapid Automatic Keyword Extraction (RAKE) algorithm, described by Jindal and Kaur [10]. This algorithm extracts key phrases from text inputs, first by dividing a document into phrases by punctuation, then into sub-phrases by various stop words, and finally into individual words by spaces. Each sequence of sub-phrases is considered a candidate keyword, and scores are assigned to every word based on the frequency of its occurrence and co-occurrence with other words in candidate keywords. The score is computed as the quotient of the total number of words in all candidate keywords containing a specific word and the word’s overall occurrence in a document.

III.c 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 can be summarized as the study of lexical semantics and the relation between sentences and paragraphs. Lexical semantics can use the characteristics of different content to classify lexical items. The task of semantic analysis is to conduct context-sensitive relation and classification reviews of data.

Tripathy et al. proposed the N-gram model, which presumes that the appearance of any given word is correlated with a selection of other words. Using a set of words with a given length, the N-gram model attempts to determine the overall contextual sentiment based on the emotions contained within. A typical implementation process would break the given text into predefined sections by word length (grams), and analyze the individual contents, before proceeding to analysis with a gram of greater size. A typical example would be to analyze the sentence "The movie is not a good one."

- Its unigram: "'The','movie','is','not','a','good','one'", would provide an overall positive result due to the presence of the word “good”.

- Its bigram: "'The movie','movie is','is not','not a','a good','good one'", which considers a pair of words at a time, would still provide the same result as the unigram.

- Its trigram: "The movie is", "movie is not ", "is not a", "not a good", "a good one", which considers three words at a time, would provide an overall *bad* result, since it would take into account the presence of “not” before “good”, and identify that as a negation.

Latent Semantic Indexing (LSI) is a very important technical idea in the field of information retrieval. LSI looks for patterns in the way words cluster together to give further background meaning to particular clusters. This clustering is done through singular value decomposition (SVD) of the term-document matrix. The basic idea behind LSI is to take advantage of implicit higher-order structure in the association of terms with documents ("semantic structure") in order to improve the detection of relevant documents, on the basis of terms found in queries [12]. LSI keywords are related to the primary keyword, providing word sense disambiguation such that “iPhone” is a keyword of "Apple", while "Apple" is a keyword of both electronic products and fruits.

Emotion recognition is an important interdisciplinary research topic in the fields of neuroscience, psychology, cognitive science, computer science, and artificial intelligence. Convolutional Neural Networks (CNN) is a statistical learning model inspired by biological neural networks, the goal of which is to automatically mark the text with defined labels. Common text classification tasks include emotion recognition, email filtering, intent identification, and data classification. Two-dimensional signals such as image and voice are hard to be modelled well by traditional models like SVM, so the ability of CNN to characterize two-dimensional signals makes it far more usable in bimodal data analysis. CNN can also adaptively extract features to eliminate the dependence on human subjectivity or experience [13].

IV. Proposed Methodology

Prior to milestone 2, the team’s plan is to develop a more coherent scoring methodology which will include the compatibility of the various models selected, as well as their dimensionality (whether one model can be used for multiple functions in the final product). The models will be selected based on their scores both according to literary review, and their performance on the selected datasets identified above. Additionally, the team will outline an approach to integrating the models, and creating the final deliverable, including either a procedural guideline outlining the necessary steps for obtaining the statistical report, or a plan to combine the models in a single executable.

Our proposed methodology consists of three key steps, starting with the identification of 9 existing models for key data and semantic analysis. These models 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:

- 3 speech recognition models

- 3 key data mining models for text data

- 3 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 80% to be used as a training set, and 20% 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.

V. Preliminary Results

From our research of prior publications, we found the following advantages and disadvantages for the models we identified for further analysis:

- Speech Recognition Models
    - The HMM model directly models the transition and performance probabilities, and counts the co-occurrence probability, which provides more usable data than the other models. There is also a lot of documentation around this model, and some of its prior implementations. However, since the state for any point in time is dependent on the previous state, the model is at a disadvantage before other models where an ad hoc approach can be used.
    - The LAS model does not make independent assumptions about the probability distribution of the output character sequence, which means the results are less affected by input bias. However, LAS requires a lot of calculations to be performed in order to convert the speech to text, and when compared to a version of the HMM, demonstrated a higher word error rate.
    - RNN models do not require prior integrations of language models, as they can operate without prior knowledge of the language of analysis. The integration process of RNN, however, is far more complex than HMM and LAS, since it requires the implementation of multiple neural networks, and their folding through backpropagation. Additionally, RNN models are heavily dependent on the pre-processing of the data being analyzed.
- Text Analysis Models
    - The TF-IDF model has a significant amount of documentation, with the code available online for perusal, and was determined through several tests to provide an accurate result, generating both topics expected based on the use of other models, and some previously unidentified within the tested dataset [8]. Additionally, while the computation requires a significant time investment, the model does not require much pre-analysis work, and can be used on any text-based dataset.
    - TextRank is widely used as a baseline for evaluating the effectiveness of keyword extraction models, due to its accuracy and high level of documentation. However, since each node can only be composed of individual words, TextRank is at a disadvantage when compared to models which provide topic mining as well as keyword extraction.

    - The RAKE model returns both words that occur more frequently and candidate keywords that have a higher coefficient of appearance and is more precise than TF-IDF and TextRank. There is significant documentation around it, as well as multiple code libraries available with various implementations in several coding languages.

- Semantic Analysis Models

    - The advantage of the N-gram model is that it contains all the information triggered by a word, but a large amount of text is required to determine the parameters of the model. The model is simple and effective, but only considers the relationship of some words, and does not consider the similarity of words, morphology, and lexology.

    - LSI, unlike the N-gram model, can partially handle the problem of polysemy - words having multiple meanings dependent on context - by adding the meanings which are relevant to the context into the analysis set. However, LSI requires recalculation and significant time investment with every reindexing of the data input.

    - The CNN model can handle high-dimensional data processing and has a good feature classification effect. However, it requires constant parameter adjustment, and a very large sample size.

Based on the initial literary analysis performed, the authors have determined an order of preference for the three models in each area of data analysis. The order was based on the criteria of levels of documentation, accessibility, and precision. The orders of preference are as follows:

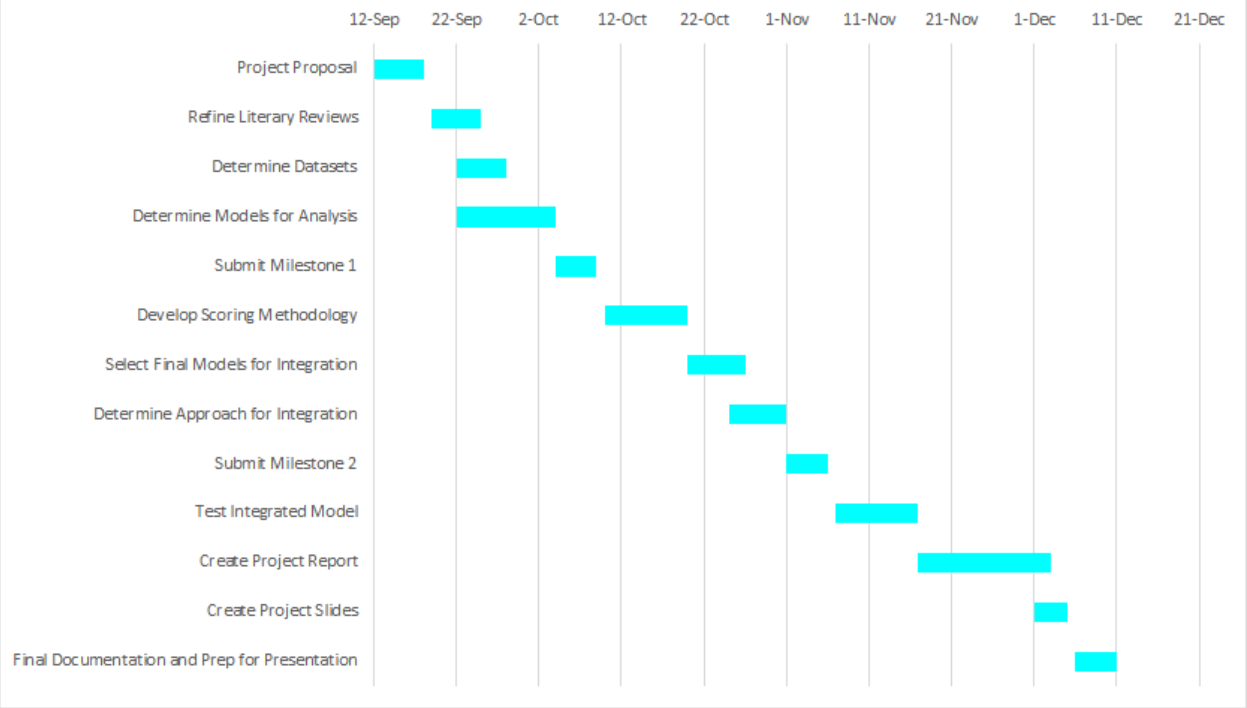
- Speech Recognition
    - Hidden Markov Models (HMM)
    - Listen, Attend, Spell (LAS)
    - Recurrent Neural Network (RNN)
- Textual keyword Mining
    - Rapid Automatic Keyword Extraction (RAKE)
    - Term Frequency-Inverse Document Frequency (TF-IDF)
    - TextRank
- Semantic Analysis
    - Latent Semantic Indexing (LSI)
    - N-gram Model
    - Convolutional Neural Networks (CNN)

Having determined these orders of preference, the authors then identified several datasets which were used to determine the practical efficiency and effectiveness of each model, and determine the selections for the final product of the research. The proposed datasets are:

- For the models involving speech recognition and speech-to-text conversion, we will use the LibriSpeech corpus, a dataset of over 1000 hours of English speech [14]. This is a commonly used dataset in the testing of audio data mining models.

- For models involving text data analysis and semantic analysis, we will use a dataset comprised of 233.1 million Amazon reviews, as this will provide a range of semantic data available for extraction [15].

V. Project Development Timeline (Gantt)



Previously we focused on identifying the aggregation model for milestone one however we have since changed this to determining all of the models that we will analyze. On our way to milestone two we will develop our scoring methodology, determine the final models and approach to integration. Lastly, after milestone two rather than scoring results we will test the integrated model and draw a conclusion of our approach.

Website Link: https://mason.gmu.edu/~acrow3/582_group_project_milestone1.html

References

[1] 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.

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

[3] Chavan, Rupali S., and Ganesh S. Sable. "An overview of speech recognition using HMM." *International Journal of Computer Science and Mobile Computing* 2.6 (2013): 233-238.

[4] W. Chan, N. Jaitly, Q. Le and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960-4964, doi: 10.1109/ICASSP.2016.7472621.

[5] Venkateswarlu, R. L. K., R. Vasantha Kumari, and G. Vani JayaSri. "Speech Recognition by Using Recurrent NeuralNetworks." *International Journal of Scientific & Engineering Research* 2.6 (2011): 1-7.

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

[7] 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.

[8] Wang, Yinying, Alex J Bowers, and David J Fikis. “Automated Text Data Mining Analysis of Five Decades of Educational Leadership Research Literature: Probabilistic Topic Modeling of EAQ Articles From 1965 to 2014.” *Educational administration quarterly* 53.2 (2017): 289–323. Web.

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

[10] Jindal, Shubhra Goyal, and Arvinder Kaur. “Automatic Keyword and Sentence-Based Text Summarization for Software Bug Reports.” *IEEE access* 8 (2020): 65352–65370. Web.

[11] Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath. “Classification of Sentiment Reviews Using n-Gram Machine Learning Approach.” *Expert systems with applications* 57 (2016): 117–126. Web.

[12] Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. “A Comparative Study of TFIDF, LSI and Multi-Words for Text Classification.” *Expert systems with applications* 38.3 (2011): 2758–2765. Web.

[13] B. Zhang, C. Quan and F. Ren, "Study on CNN in the recognition of emotion in audio and images," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 2016, pp. 1 -5, doi: 10.1109/ICIS.2016.7550778

[14] Panayotov, Vassil, et al. "Librispeech: an asr corpus based on public domain audio books." 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015.

[15] Ni, Jianmo, Jiacheng Li, and Julian McAuley. "Justifying recommendations using distantly-labeled reviews and fine-grained aspects." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.