A Research Pathway for Stance Detection in Speech

Vy Nguyen
School of Science, Engineering & Tech
RMIT University Vietnam
Ho Chi Minh City, Vietnam
s3964786@rmit.edu.vn

Xiuzhen Zhang
School of Computing Technologies
RMIT University
Melbourne, VIC, Australia
xiuzhen.zhang@rmit.edu.au

Andrew M. Colarik
School of Science, Engineering & Tech
RMIT University Vietnam
Ho Chi Minh City, Vietnam
andrew.colarik@rmit.edu.vn

***Abstract*—Stance detection is a Natural Language Processing (NLP) task that involves identifying an individual's standpoint on a specific topic and determining their stance as either in favor of or against it. It has various potential applications for governments and organizations, helping them understand how public opinion evolves across multiple social media platforms. Although humans use both written and spoken language for communication, most research in the field primarily focuses on text-based data for building models. This paper aims to put forth a novel research pathway for developing a stance detection model that is specifically trained on audio-based data. To establish a direction for our approach, we analyze several foundational factors, including the utterance of the stance and the emotional elements present in human speech and conversations. We subsequently construct a flow to process the data by extracting text- and audio-based features, as well as supplementary attributes such as the author's profile. A two-phase modeling approach is proposed to integrate text and speech into a single ensemble model, aiming to enhance the accuracy of the predictions. Finally, several improvement opportunities are also identified, providing a baseline for future endeavors in audio-based stance detection.**

Keywords—stance, stance detection, speech analysis, opinion mining, text mining

I. Introduction

Nowadays, content creation enables individuals to communicate their viewpoints on various controversial subjects and socio-political events [1]. There has been a growing trend of individuals utilizing multimedia platforms to share, discuss, and exchange opinions [2]. As a result, recently, users of leading technology companies such as Twitter, TikTok, Spotify, and Meta have been actively contributing content across all domains and topics. This phenomenon gives rise to a vast repository of underexploited data generated by internet users [3]. The huge dependency of internet users on these social network platforms as a primary source of communication allows researchers to study different aspects of online human behavior, including the public's stance toward various social and political matters.

Understanding the diverse perspectives, or stances, expressed in user-generated content (UGC) is increasingly important for governments, organizations, and individuals [4]. In NLP, the automation of this task is called *stance detection* [4, 7, 15, 17]. It aims to identify an individual's position on a specific topic to determine whether they support or oppose it [5, 17]. In recent years, societal subjects such as racism, global warming, and feminism have been frequently cited as themes for stance detection on social media [6]. Similarly, political subjects, including referendums and elections, have been extensively studied in stance detection research to analyze public opinion [6]. Stance detection is employed in these contexts because it facilitates a comprehensive understanding of UCG, extending beyond its literal interpretation. The applications of this encompass social media impact analysis, rumor evaluation, misinformation mitigation, targeted advertising, and temporal evolution analysis of public opinions [7]. This makes stance detection a significant problem to solve.

While internet users use both written and spoken language to produce content on social media platforms, stance detection in spoken language is still an understudied research area [7]. Existing stance detection approaches are primarily trained on textual corpora [7], hence they often overlook paralinguistic and non-linguistic cues in human speech. In human speech, paralinguistic information includes intention, attitude, and style, while non-linguistic information encompasses sentiment and emotion [8], all of which can be an indicator of the speaker's stance. Humans can infer various pieces of information in oral communication through cues such as pitch, pace, and intensity. Previous research has attempted to convert speech into text for the purpose of analyzing stance [9]. However, NLP models built solely on transcription have been unsuccessful in capturing emotional elements such as laughter or variations in volume and have also faced challenges with inaccuracies in the transcription models [10]. Therefore, a novel model capable of accurately detecting stance by effectively incorporating both text and speech modalities would be revolutionary in this research field.

In this paper, we will propose a stance detection approach that integrates text and speech modalities to improve detection accuracy. The paper initially provides a formal definition of the stance detection task in speech. Next, it explores the foundational aspects of stance detection and provides an overview of how existing studies have addressed the challenges of understanding human language in both spoken and written forms. It then proceeds to outline a new research pathway that combines text- and audio-based features to identify stance and provides an overview of the methodology for an experimental setup. The discussion will primarily focus on several potential techniques that can be selected for constructing the model and evaluating its performance.

II. Problem Definition

Before formally defining the problem, it is advantageous to provide a linguistic definition of stance. Du Bois defines stance as "*a public act by a social actor, achieved dialogically through overt communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects with respect to any salient dimension of the sociocultural field*" [11]. In simple terms, during the communication process, individuals tend to develop and express their stance towards a target of interest in search of alignment with other individuals. For instance, given the target "*global warming is a real concern*", the tweet "*The biggest threat in this world is global warming #floods #icemelting*", although conveys anger, suggests that the author potentially supports the topic. Stance detection aims to train machines to automatically make this inference.

Stance detection, also known as *stance identification*, *stance analysis*, or *attitude detection*, refers to "*the task of automatically determining whether the author of a text is in favor of, against, or neutral toward a proposition or a target*"

[12]. In NLP, stance is typically represented using discrete class labels, namely positive or favorable, negative or unfavorable, and neutral or none [12]. Vychegzhanin et al. [13] formulates the task as follows:

> *For the given corpus [...] D containing the authors' points of view on the target object g, and the stance scale S, construct a function (classifier) c:*
>
> $$c: D(g) \times S \rightarrow \{true, false\}$$
>
> *The stance scale S can be binary (favor-against), ternary (favor-against-neither), or n-ary.*

Simply put, the classifier function *c* determines whether the content belongs to each of the given classes of the author's opinions. In practice, a variety of entities, including people, products, services, political parties, companies, and social movements, can serve as target objects for the classifier [13]. The problem definition remains unchanged in the context of stance analysis in speech. Only the data is now audio-based.

III. Foundational Considerations

Existing stance detection models primarily focus on text-based data [7], which implies that its applications are currently restricted to analyzing textual or transcribed user-generated content only. As a result, the ability to analyze stance in spoken language will allow for the expansion of these applications across different multimedia channels and platforms. This section will explore the fundamental aspects of detecting stance in speech. These include the understanding of stance utterance, the relationship with other tasks in opinion mining, the extraction of features from text and speech, the modeling at user level, and the limitations of current evaluation methods.

A. Stance Utterance

Understanding stance utterance is a key component in the task of identifying stance. Stance utterance pertains to the stance conveyed through a written or spoken message [14]. It comprises the constituent elements of a particular viewpoint and aids in the interpretation of the message by others [14]. Determining stance utterance is a complex task that requires accurately identifying the intended target, which can be explicitly stated, implied, indirectly referenced, or partially mentioned [15]. Subtle language nuances in spoken language can also make it challenging to determine the speaker's intention [16]. Various methods exist for individuals to express their opinions, including implicit, explicit, ironic, metaphorical, or uncertain expressions [15]. Some examples of this phenomenon in verbal communication include the utilization of negating words, variations in pitch, pauses, and laughter [16]. These linguistic features can potentially alter the literal meaning of a piece of content, making its actual meaning highly tricky to determine. Thus, identifying an individual's stance in verbal communication requires a comprehensive understanding of the nuances present in spoken language.

B. Sentiment and Sarcasm

Stance detection is closely linked to sentiment analysis and sarcasm detection [17]. While stance detection is concerned with identifying the speaker's perspective, sentiment analysis classifies the emotional polarity of the content as positive, negative, or neutral [18]. Compared to sentiment analysis, stance detection considers various additional factors, including the content-author relationship, the target(s), and the context, which may not be readily apparent [15]. Sarcasm, on the other hand, is a linguistic phenomenon that may completely reverse one's stance [19]. The presence of sarcasm has been found to contribute to inaccuracies in text-based stance detection models [20]. Razali et al. point out that this linguistic phenomenon greatly complicates the analysis of spoken language [21]. Given the interconnected nature of these NLP tasks, it is impractical to approach stance detection without considering previous research on sentiment analysis and sarcasm detection. In fact, the adaptation of existing pretrained models on these related tasks has the potential to enhance the performance of the stance detection task.

C. Feature Extraction

To effectively model the problem, it is necessary to extract appropriate features from the content. According to Alturayeif et al., the majority of stance detection studies focus on content-level modeling [7]. This approach involves extracting five feature clusters: semantic, syntactic, structural, statistical, and pragmatic [7]. Vychegzhanin et al. propose that higher-level linguistic features, including target-indicative, stance-indicative, stylistic, and sentiment features, are also valuable for the task as well [13]. Furthermore, NLP tasks involving knowledge acquisition from audio-based data typically take into account audio features. Previous studies have utilized various variables such as time, frequency, amplitude, spectral energy, and the raw waveform of the sound for other NLP tasks such as sentiment analysis [22] and sarcasm detection [23] in speech. These audio features can be used as a basis for constructing other higher-level linguistic features mentioned before [24]. For example, a rise in volume may indicate heightened emotional arousal; a pause may suggest reluctance; and speed acceleration can imply embarrassment. Even though these audio features could provide more parameters for modeling, it is important to note that NLP models that use low-level features might be more likely to overfit onto irrelevant signals [25]. Therefore, it is essential to conduct a feature importance analysis for linguistic and audio features in order to determine the appropriate features for the final model.

D. User-level Modelling

Stance detection can be modeled at the user level in addition to content-level feature extraction [7]. The author's background and temporal stance evolution are significant factors in their profile that impact their current stance towards a target of interest [11]. The author's background, including occupation, demographic information, religion, and ideology, usually suggests a stance towards a topic [26]. On the other hand, one's stance can evolve over time due to influence from media platforms that are actively producing a significant amount of content across various channels [27]. To both capture content-level features and user-level features, Chen and Ku propose a hybrid technique that integrates on-topic content with the author's interaction network, preference network, and connection network [28]. This approach leads to a significant improvement in the performance of the model. Figure 1, presented by Alturayeif et al. [7], provides a summary of the feature extraction techniques employed in the hybrid approach.

In general, research involving author-related features has the potential to compromise user privacy [29]. Social media platforms have implemented several privacy measures to

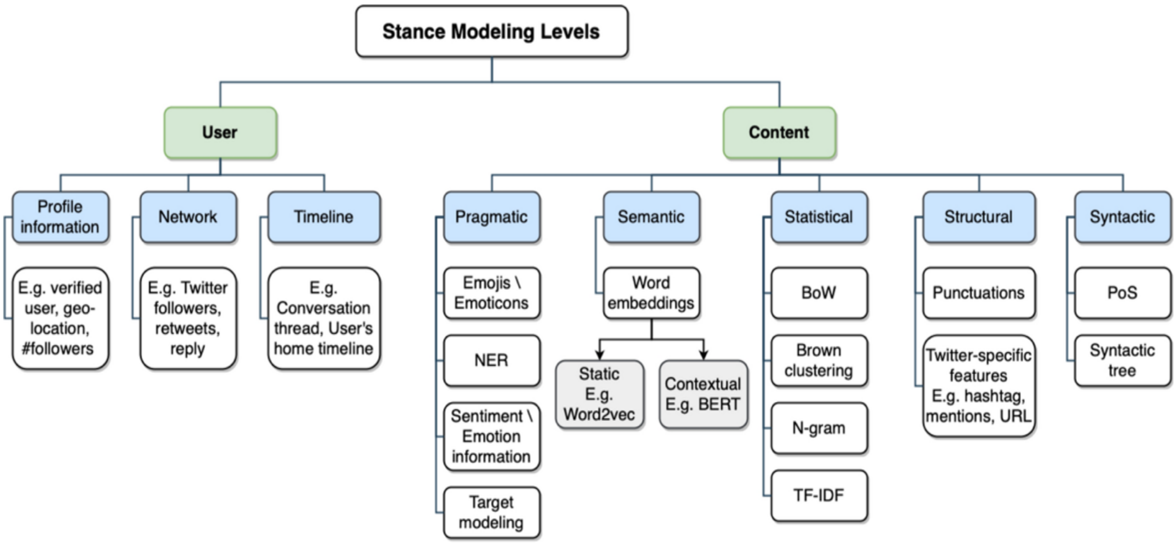

Figure 1. Content-level and user-level feature extraction

restrict access to user information [30]. This emphasizes the necessity for additional effort to safeguard social media users from inadvertently revealing their personal information and opinions through the model, even with their consent.

E. Challenges in model evaluations

Effective model evaluation methods are crucial to optimizing model performance. However, current approaches to stance detection, including traditional machine learning (ML) models, deep learning architectures, and transfer learning models, all have limitations during the evaluation phase [7]. The primary contributing factor to inefficient evaluation is the lack of quality annotated training data [17]. The data annotation process can be challenging when there are multiple targets or when the author's viewpoint on a specific target changes over time [31]. A shift in stance can also happen when a concept changes its polarity across different domains, across languages, and across geographies [32, 33]. Furthermore, the taskers' own biases can impact the labels they assign during data annotation, thereby increasing the complexity of the task [34]. As a result, numerous models exhibit strong performance on trained data but demonstrate subpar performance in real-world scenarios [15]. This is particularly evident when these models are applied to unfamiliar data or domains, lack contextual information, or are utilized in time periods different from those they were trained on [15]. Due to these dynamics of human language, it is therefore necessary to regularly fine-tune and retrain models to minimize performance degradation. In certain use cases, domain-specific or organization-specific models might need to be considered to achieve better control over annotated data.

IV. Research Approach

This section proposes a new research approach that can potentially overcome the limitations of existing techniques for detecting stance. The proposed approach integrates both text-based and audio-based features. We will use the Spotify Podcast Dataset, containing 100,000 podcasts [35], to develop a two-phase model that effectively identifies stance by considering multiple dimensions. During Phase 1, relevant features will be extracted from pre-processed data and further enhanced into higher-level features, such as sentiment, arousal, and sarcasm features. Phase 2 involves clustering data to create distinct groups based on domains, topics, and authors. After that, an ensemble classifier will be introduced to assign a stance label to each podcast. A multi-step evaluation strategy will be proposed to address the current limitations of model evaluations. Figure 2 depicts the overall architecture of our proposed two-phase approach, which will be described in detail subsequently.

A. The Spotify Podcast Dataset

In 2020, Spotify launched the Spotify Podcast Dataset, which consists of 100,000 podcasts with approximately 60,000 hours of speech [35]. The podcasts encompass various subjects such as cultural and lifestyle storytelling, sports, news, wellness, documentaries, and politics [35]. Labeled data for retrieval and summarization is generated for each podcast in the corpus [35]. English is the predominant language used, with a number of podcasts available in multiple languages [35]. Spotify also utilized Google's Cloud Speech-to-Text API [1] to automatically generate text transcripts [35]. The metadata accompanying each podcast, such as author profiles and timelines, automatic labels, summaries, and transcriptions, serves as a valuable foundation for the subsequent steps.

We propose utilizing this dataset for our research for three reasons. First, the corpus is claimed to be the largest collection of transcribed speech data [35]. Second, it exhibits a significant range of domains, topics, and speaker backgrounds [35]. Third, the corpus is being expanded to include Spanish and other languages [35], allowing for the potential extension of our model for detecting stance in different languages and multilingual contexts. It is worth mentioning that the transcription result provided, with a 81.8% sample named entity recognition accuracy and an 18.1% word error rate [35], is not as accurate as the state-of-the-art results on other corpora. For example, the word error rate on Switchboard [36] is less than 5% [37]. This high error rate might be attributed to the impromptu and conversational nature of podcasts.

[1] https://cloud.google.com/speech-to-text

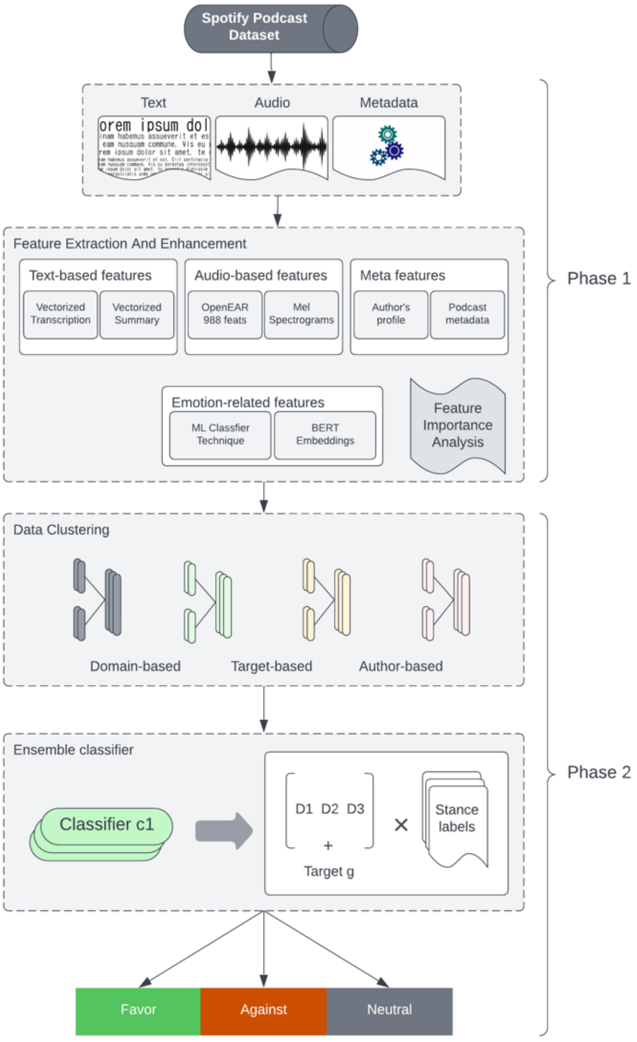

Figure 2. Overall architecture of the two-phase stance detection model

B. *Phase 1: Feature Engineering and Enhancement*

Given the Spotify Podcast Dataset, we will start the modeling process by examining transcribed texts. For this step, we will use traditional NLP techniques, such as *n-gram* features, tokenization, lemmatization, and vectorization [38], to create a vectorized representation of the transcribed podcasts. An exploratory data analysis of the transcription should be conducted before selecting the appropriate NLP techniques to be used for modeling. Nevertheless, it is worth considering vectorizing the summary for each podcast as well. This will be particularly beneficial for longer podcasts, those that have an impromptu style, and those with multiple speakers, as the summary can potentially provide a clearer indication of the stance compared to the entire content [39]. Author information, if available, should also be gathered at this stage to enable data clustering by author and subsequent analysis of their evolving stance.

For audio-based feature extraction, we propose utilizing two candidates to extract the data in two different ways: (1) OpenEAR and (2) Mel spectrograms. OpenEAR, or Open-source Emotion and Affect Recognition, is open-source software designed for extracting audio features and optimized for recognizing emotions and affect in speech [40]. OpenEAR is capable of extracting 988 audible features that are appropriate for sentiment analysis [40]. These features include low-level descriptors (LLDs) such as intensity, loudness, pitch, probability of voicing, and the regression coefficients derived from the LLDs [40]. The features extracted using openEAR provide an extensive emotional vector representation for an audio segment. This representation can be utilized for emotion, sentiment, and stance classification tasks. Mel spectrograms, on the other hand, provide a Mel scale, which is a logarithmic transformation of an audio signal's frequency [41]. The Mel scale is commonly used in audio processing related to human perception due to its alignment with how humans perceive sound [42]. The reason is that sounds that have the same Mel scale distance are perceived as equivalent by human ears [43]. Therefore, the Mel scale is also a good candidate for modeling higher-level emotion-related features.

After extracting fundamental audio and text-based features, we can proceed to derive additional features that describe higher-level characteristics of the content, such as the sentiment, arousal, and emotion, from them. OpenEAR discrete features can be fed into conventional machine learning classifiers to determine discrete labels that describe the author's sentiment and emotions [44]. Simultaneously, the Mel scale generated by Mel spectrograms, can be used well with neural classifiers like sequence-CNN (convolutional neural network) to label the data [45]. For extracted text-based features, lexicon-based methods can be utilized to determine relevant labels. This conventional NLP technique employs an emotion lexicon, such as the NRC Emotion Lexicon[2] and the NRC Emotion Intensity Lexicon[3], in order to assign an appropriate an label and measure the emotional intensity of the transcribed document.

To mitigate the contextual loss resulting from the conventional NLP processing techniques, pre-trained word embedding models that capture contextualized semantics can be employed. BERT, or Bidirectional Encoder Representations from Transformers, developed by Google, can be considered to extract the emotion-related features as well [46]. Research has demonstrated that BERT [CLS] and sentence-BERT embeddings are capable of effectively capturing contextual information [47], suggesting their potential for improved annotation of emotion-related features. It must be noted that, all the mentioned approaches are performed against different sets of extracted features, consequently, they have the potential to produce varying or even conflicting labels. Therefore, we should finally employ a feature selection technique such as weighted average or majority voting to decide the best labels for a podcasts.

C. *Phase 2: Stance Detection Modeling*

It is pointed out earlier that performance degradation in stance detection usually takes place due to model overfitting on the training data. To tackle this issue, we will partition the dataset into distinct clusters and subsequently conduct training and validation procedures for each cluster individually as well as for the dataset as a whole. The clustering criteria can be categorized into three types: domain-based, topic-based, and author-based. The metadata accompanying each podcast in the dataset can help determine these categories. The utilization of domain-based and topic-based groups can facilitate the

[2] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[3] https://saifmohammad.com/WebPages/AffectIntensity.htm

acquisition of domain-specific knowledge and mitigate potential semantic ambiguities that may arise when dealing with diverse disciplinary contexts. Author-based groups, on the other hand, can serve as a valuable means to identify and analyze temporal shifts in an individual's stance on a given topic [48]. By examining the collective body of podcasts within the dataset, these groups can contribute to a comprehensive understanding of the interconnections and dynamics between them.

In the last step of the training, an ensemble classifier will be introduced. This classifier utilizes all the extracted features that represent a document in order to classify it against the target. The training will yield a generic classifier and multiple cluster-specific classifiers. This use of two parallel training pipelines help highlight prediction discrepancies between a cluster-specific model and a generic model, providing valuable insights for optimizing the final model. Additionally, the use of narrow domain adaptive models can address the problem of catastrophic forgetting in continuous learning as each unique domain, topic, and author has its own adapter [49]. The labels for the final classification are Favor, Against, and Neutral, which can be unified from all the individual classifiers using voting techniques.

D. *Evaluation Strategy*

Research on information retrieval and information extraction frequently employs the metrics of precision, recall, and F-score (or F-measure) [17]. F-score (F) is a combined metric that is calculated using precision (P) and recall (R), with the option to assign weights to these two metrics [50]. However, given the subjective nature of stance, the reliability of the data annotation process remains crucial for these metrics to capture an accurate reflection of the reality. A shift in stance can occur when the same concept alters its polarity across different domains [51]. For instance, high prices may indicate an unfavorable stance in the consumer market and also a favorable stance in the stock market. A semantic shift can also occur across languages and cultures [52]. A pitch raise, for example, might potentially convey an opposing stance in English, whereas in Vietnamese and Chinese, it may solely stem from the tonal characteristics of the language. Consequently, if we solely depend on annotated data for refining the model and optimizing the metrics, there remains a risk of a performance degradation when the model is applied in real-world scenarios. It is therefore necessary for our evaluation to address these biases effectively.

We hereby propose a multi-step evaluation strategy to assess the model's performance that consider various factors that might contribute to performance drift. Due to the resource-intensive nature of audio data processing, our initial step will involve sampling podcasts pertaining to various domains, topics, and authors, followed by the annotation of this data. We propose to sample only 10% to 20% of the dataset size and reserve the rest for later steps, however, this step should prioritize the inclusion of diverse data samples to prevent the model from performing poorly on unfamiliar domains later. The labeled data can be divided into a training set and a test set, with the test set remaining untouched during model training. The test set is then used to score the model using the above metrics and help finetune the training process.

After achieving satisfactory performance on labeled data, we will proceed to apply the model to unlabeled data. The outcomes of this step can be applied in two directions. In the first direction, we select the labels with the highest confidence score and utilize them as pseudo-labeled data for the purpose of retraining the model. Semi-supervised machine learning employs this technique to address the limited availability of labeled data [53]. In the second direction, human involvement is incorporated to obtain feedback on the classified labels. It will be ideal if we can involve end users of the model in this stage, as they will provide early feedback based on their interpretations of the stance labels and their expectations of the model performance. This feedback can be subsequently utilized as new labeled data to enhance the model as well. It is beneficial to develop an iterative feedback pipeline that automatically incorporates new labeled data, including pseudo labels and end user feedback, and consistently refines the model. Figure 3 provides an overview of our evaluation approach for better visualization.

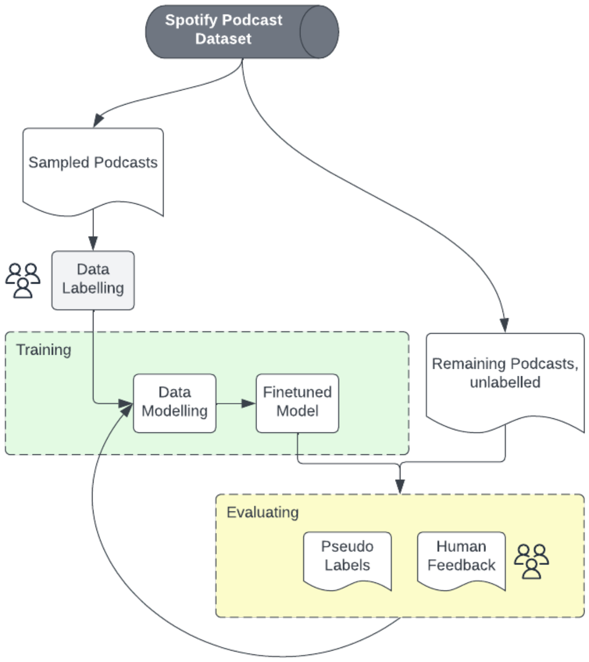

Figure 3. Model evaluation iterative approach

In order to conduct a comprehensive evaluation, we also propose to compare the optimized model obtained in the final stage with existing works on stance detection. Acknowledging a lack of research on stance detection in spoken language, one possible strategy is to apply pre-existing models to transcribed podcast data and subsequently apply our own model to the original podcasts. The disparity in performance between the two executions will provide insights into how audio data can influence the analysis of stance in human language.

V. Conclusion

This paper examines the topic of stance detection in speech, analyzes important aspects of the problem, and proposes a novel research approach to address the research gap of previous approaches. It then introduces a two-phase strategy. This two-phase strategy enhances data modeling by incorporating both text-based and audio-based features, enabling us to capture more information compared to previous models. Furthermore, high-level emotion-related labels are extracted from both modalities and contextual loss prevention measures are also considered. The data is organized into clusters based on domains, topics, and authors. This allows the ensemble model to effectively capture specific characteristics and knowledge within each cluster. Our evaluation strategy proposes a semi-supervised approach that incorporates human involvement in an iterative retraining pipeline for ongoing model refinement.

That being said, there are several directions to extend the proposed research. The study utilizes the Spotify Podcast Dataset, which is currently limited to English and Spanish. Hence, it is necessary to investigate additional datasets for training multi-lingual and cross-lingual models. Also, it is indicated that automatic transcriptions in the dataset exhibit a notable error rate, which potentially compromises the quality of features extracted from transcribed data. This presents an opportunity to enhance the model's performance through the enhancement of the transcription process. On the other hand, our approach requires manual data annotation to train the initial iterations. The process can be labor-intensive and challenging, especially when dealing with audio data from various domains, topics, and timelines. Furthermore, processing audio data is generally resource-intensive, so it is essential to thoroughly analyze computational feasibility when developing the iterative retraining pipeline. Given these areas for improvement, we believe that the findings presented in this paper will serve as a solid foundation for future research on speech-based stance detection.

REFERENCES

[1] S. Stieglitz and L. Dang-Xuan, "Social media and political communication: a social media analytics framework," *Soc. Netw. Anal. Min.*, vol. 3, no. 4, pp. 1277–1291, Dec. 2013, doi: 10.1007/s13278-012-0079-3.

[2] Y. A. Ahmed, M. N. Ahmad, N. Ahmad, and N. H. Zakaria, "Social media for knowledge-sharing: A systematic literature review," *Telematics and Informatics*, vol. 37, pp. 72–112, Apr. 2019, doi: 10.1016/j.tele.2018.01.015.

[3] M. Blackburn, J. Alexander, J. D. Legan, and D. Klabjan, "Big Data and the Future of R&D Management," *Research-Technology Management*, vol. 60, no. 5, pp. 43–51, Sep. 2017, doi: 10.1080/08956308.2017.1348135.

[4] S. Ghosh, P. Singhania, S. Singh, K. Rudra, and S. Ghosh, "Stance Detection in Web and Social Media: A Comparative Study," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 75–87. doi: 10.1007/978-3-030-28577-7_4.

[5] F. Alqasemi, H. Al-Baadani, and M. A. Al-Hagery, "Stance Detection using Two Popular Benchmarks: A Survey," in *2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, Oct. 2022, pp. 1–6. doi: 10.1109/eSmarTA56775.2022.9935138.

[6] A. ALDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Information Processing & Management*, vol. 58, no. 4, p. 102597, Jul. 2021, doi: 10.1016/j.ipm.2021.102597.

[7] N. Alturayeif, H. Luqman, and M. Ahmed, "A systematic review of machine learning techniques for stance detection and its applications," *Neural Comput & Applic*, vol. 35, no. 7, pp. 5113–5144, Mar. 2023, doi: 10.1007/s00521-023-08285-7.

[8] B. T. Atmaja and A. Sasou, "Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations," *Sensors*, vol. 22, no. 17, Art. no. 17, Jan. 2022, doi: 10.3390/s22176369.

[9] G.-A. Levow *et al.*, "Recognition of stance strength and polarity in spontaneous speech," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2014, pp. 236–241. doi: 10.1109/SLT.2014.7078580.

[10] N. Kaur and P. Singh, "Conventional and contemporary approaches used in text to speech synthesis: a review," *Artif Intell Rev*, vol. 56, no. 7, pp. 5837–5880, Jul. 2023, doi: 10.1007/s10462-022-10315-0.

[11] D. Bois and J. W, "The stance triangle," in *Stancetaking in Discourse: Subjectivity, evaluation, interaction*, R. Englebretson, Ed., in Pragmatics & Beyond New Series. John Benjamins Publishing Company, 2007, pp. 139–182. doi: 10.1075/pbns.164.07du.

[12] M. Lai, A. T. Cignarella, D. I. Hernández Farías, C. Bosco, V. Patti, and P. Rosso, "Multilingual stance detection in social media political debates," *Computer Speech & Language*, vol. 63, p. 101075, Sep. 2020, doi: 10.1016/j.csl.2020.101075.

[13] S. Vychegzhanin and E. Kotelnikov, "A New Method for Stance Detection Based on Feature Selection Techniques and Ensembles of Classifiers," *IEEE Access*, vol. 9, pp. 134899–134915, 2021, doi: 10.1109/ACCESS.2021.3116657.

[14] X. Shi, J. Wu, and L. Wei, "Stance Taking in News Interviews Based on Stance Triangle and Conversational Analysis," *Open Journal of Modern Linguistics*, vol. 12, no. 2, Art. no. 2, Mar. 2022, doi: 10.4236/ojml.2022.122015.

[15] R. Alkhalifa and A. Zubiaga, "Capturing stance dynamics in social media: open challenges and research directions," *Int J Digit Humanities*, vol. 3, no. 1, pp. 115–135, Apr. 2022, doi: 10.1007/s42803-022-00043-w.

[16] C. Davies, V. Porretta, K. Koleva, and E. Klepousniotou, "Speaker-Specific Cues Influence Semantic Disambiguation," *J Psycholinguist Res*, vol. 51, no. 5, pp. 933–955, Oct. 2022, doi: 10.1007/s10936-022-09852-0.

[17] D. Küçük and F. Can, "Stance Detection: A Survey," *ACM Comput. Surv.*, vol. 53, no. 1, p. 12:1-12:37, Feb. 2020, doi: 10.1145/3369026.

[18] D. S. Chauhan, R. Kumar, and A. Ekbal, "Attention Based Shared Representation for Multi-task Stance Detection and Sentiment Analysis," in *Neural Information Processing*, T. Gedeon, K. W. Wong, and M. Lee, Eds., in Communications in Computer and Information Science. Cham: Springer International Publishing, 2019, pp. 661–669. doi: 10.1007/978-3-030-36802-9_70.

[19] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic Sarcasm Detection: A Survey," *ACM Comput. Surv.*, vol. 50, no. 5, p. 73:1-73:22, Sep. 2017, doi: 10.1145/3124420.

[20] S. Ghosh, P. Singhania, S. Singh, K. Rudra, and S. Ghosh, "Stance Detection in Web and Social Media: A Comparative Study," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 75–87. doi: 10.1007/978-3-030-28577-7_4.

[21] M. S. Razali, A. A. Halin, N. M. Norowi, and S. C. Doraisamy, "The importance of multimodality in sarcasm detection for sentiment analysis," in *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, Dec. 2017, pp. 56–60. doi: 10.1109/SCORED.2017.8305421.

[22] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, "Speech Sentiment Analysis via Pre-Trained Features from End-to-End ASR Models," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7149–7153. doi: 10.1109/ICASSP40776.2020.9052937.

[23] A. Mathur, V. Saxena, and S. K. Singh, "Understanding sarcasm in speech using mel-frequency cepstral coefficent," in *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, Jan. 2017, pp. 728–732. doi: 10.1109/CONFLUENCE.2017.7943246.

[24] O. Lahaie, R. Lefebvre, and P. Gournay, "Influence of audio bandwidth on speech emotion recognition by human subjects," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov. 2017, pp. 61–65. doi: 10.1109/GlobalSIP.2017.8308604.

[25] H. Abburi, M. Shrivastava, and S. V. Gangashetty, "Improved multimodal sentiment detection using stressed regions of audio," in *2016 IEEE Region 10 Conference (TENCON)*, Nov. 2016, pp. 2834–2837. doi: 10.1109/TENCON.2016.7848560.

[26] K. Darwish, W. Magdy, and T. Zanouda, "Improved Stance Prediction in a User Similarity Feature Space," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, in ASONAM '17. New York, NY, USA: Association for Computing Machinery, Jul. 2017, pp. 145–148. doi: 10.1145/3110025.3110112.

[27] K. S. Hasan and V. Ng, "Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 751–762. doi: 10.3115/v1/D14-1083.

[28] W.-F. Chen and L.-W. Ku, "UTCNN: a Deep Learning Model of Stance Classification on Social Media Text," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1635–1645. Accessed: Sep. 25, 2023. [Online]. Available: https://aclanthology.org/C16-1154

[29] D. Dogan, B. Altun, M. S. Zengin, M. Kutlu, and T. Elsayed, "Catch Me If You Can: Deceiving Stance Detection and Geotagging Models to Protect Privacy of Individuals on Twitter," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, pp. 173–184, Jun. 2023, doi: 10.1609/icwsm.v17i1.22136.

[30] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, in WPES '05. New York, NY, USA: Association for Computing Machinery, Nov. 2005, pp. 71–80. doi: 10.1145/1102199.1102214.

[31] B. Schiller, J. Daxenberger, and I. Gurevych, "Stance Detection Benchmark: How Robust is Your Stance Detection?," *Künstl Intell*, vol. 35, no. 3, pp. 329–341, Nov. 2021, doi: 10.1007/s13218-021-00714-w.

[32] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, "Diachronic word embeddings and semantic shifts: a survey," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1384–1397. Accessed: Sep. 26, 2023. [Online]. Available: https://aclanthology.org/C18-1117

[33] Q. Sun, Z. Wang, Q. Zhu, and G. Zhou, "Exploring Various Linguistic Features for Stance Detection," in *Natural Language Understanding and Intelligent Applications*, C.-Y. Lin, N. Xue, D. Zhao, X. Huang, and Y. Feng, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 840–847. doi: 10.1007/978-3-319-50496-4_76.

[34] S. Peng, Y. Wang, D. Yu, and P. Liu, "Perception and Cognition Matters: A new light on sentiment analysis task," in *2022 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2022, pp. 1–7. doi: 10.1109/IJCNN55064.2022.9892719.

[35] A. Clifton *et al.*, "100,000 Podcasts: A Spoken English Document Corpus," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5903–5917. doi: 10.18653/v1/2020.coling-main.519.

[36] "Switchboard-1 Release 2 - Linguistic Data Consortium." https://catalog.ldc.upenn.edu/LDC97S62 (accessed Sep. 26, 2023).

[37] W. Wang, G. Wang, A. Bhatnagar, Y. Zhou, C. Xiong, and R. Socher, "An investigation of phone-based subword units for end-to-end speech recognition." arXiv, Jun. 21, 2021. doi: 10.48550/arXiv.2004.04290.

[38] M. T. H. K. Tusar and M. T. Islam, "A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data." arXiv, Oct. 02, 2021. doi: 10.48550/arXiv.2110.00859.

[39] R. Sepúlveda-Torres, M. Vicente, E. Saquete, E. Lloret, and M. Palomar, "Exploring Summarization to Enhance Headline Stance Detection," in *Natural Language Processing and Information Systems*, E. Métais, F. Meziane, H. Horacek, and E. Kapetanios, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 243–254. doi: 10.1007/978-3-030-80599-9_22.

[40] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR — Introducing the munich open-source emotion and affect recognition toolkit," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, Sep. 2009, pp. 1–6. doi: 10.1109/ACII.2009.5349350.

[41] S. Ueno and T. Kawahara, "Phone-Informed Refinement of Synthesized Mel Spectrogram for Data Augmentation in Speech Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 8572–8576. doi: 10.1109/ICASSP43922.2022.9746582.

[42] S. Vasquez and M. Lewis, "MelNet: A Generative Model for Audio in the Frequency Domain." arXiv, Jun. 04, 2019. doi: 10.48550/arXiv.1906.01083.

[43] M. Zakariah, R. B, Y. Ajmi Alotaibi, Y. Guo, K. Tran-Trung, and M. M. Elahi, "An Analytical Study of Speech Pathology Detection Based on MFCC and Deep Neural Networks," *Comput Math Methods Med*, vol. 2022, p. 7814952, Apr. 2022, doi: 10.1155/2022/7814952.

[44] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, Jan. 2016, doi: 10.1016/j.neucom.2015.01.095.

[45] J. Li, X. Zhang, L. Huang, F. Li, S. Duan, and Y. Sun, "Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neutral Network," *Applied Sciences*, vol. 12, no. 19, Art. no. 19, Jan. 2022, doi: 10.3390/app12199518.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.

[47] L. Huang, C. Sun, X. Qiu, and X. Huang, "GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3509–3514. doi: 10.18653/v1/D19-1355.

[48] A. Almadan, M. L. Maher, and J. Windett, "Stance Detection for Gauging Public Opinion: A Statistical Analysis of the Difference Between Tweet-Based and User-Based Stance in Twitter," in *Advances in Information and Communication*, K. Arai, Ed., in Lecture Notes in Networks and Systems. Cham: Springer Nature Switzerland, 2023, pp. 358–374. doi: 10.1007/978-3-031-28076-4_27.

[49] Y. Xu, X. Zhong, A. J. J. Yepes, and J. H. Lau, "Forget Me Not: Reducing Catastrophic Forgetting for Domain Adaptation in Reading Comprehension," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9206891.

[50] D. C. Blair, "Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp. Price: $32.50," *Journal of the American Society for Information Science*, vol. 30, no. 6, pp. 374–375, 1979.

[51] S. Wang, G. Lv, S. Mazumder, and B. Liu, "Detecting Domain Polarity-Changes of Words in a Sentiment Lexicon." arXiv, Apr. 29, 2020. doi: 10.48550/arXiv.2004.14357.

[52] J. C. Jackson *et al.*, "Emotion semantics show both cultural variation and universal structure," *Science*, vol. 366, no. 6472, pp. 1517–1522, Dec. 2019, doi: 10.1126/science.aaw8160.

[53] Y. C. A. P. Reddy, P. Viswanath, and B. E. Reddy, "Semi-supervised learning: a brief review," *International Journal of Engineering & Technology*, vol. 7, no. 1.8, Art. no. 1.8, Feb. 2018, doi: 10.14419/ijet.v7i1.8.9977.