Multi-Label Movie Genre Classification from Plot Summaries

Akhil Patel, Sharad Varadarajan
University of California, Berkeley, CA, School of Information
{asp82, sharadv}@berkeley.edu

Abstract-- **In this study, we perform multi-label, multi-class movie genre classification on plot summaries from a considerably imbalanced dataset. Specifically, we explore different ways to construct informative plot summary vectors, including through the use of a novel integration of Universal Sentence Encodings (USE) with Hierarchical Attention Networks (HAN). We also address problems faced by past researchers who used this corpus by combining multi-label sampling with homogeneous ensemble methods to effectively train models on de-skewed data. Each exploratory ensemble of classifiers is compared to a counterpart trained on the full imbalanced data, as well as to a bag-of-words TFIDF baseline. We find that utilization of USE and HAN can outperform the baseline though improvements, as measured by precision, recall, and F1-score, are incremental at best.**

I. Introduction

In today's world, consumers have access to many sources of information about movies they may wish to see. From movie trailers to plot summaries to reviews by critics and fellow users, it is typical for a consumer to take some of this information into account before making his or her selection. Genre tends to be a primary factor users consider when making their selection. For example, a consumer is in the mood for either a horror film or a comedy but usually not both. When reading a plot summary, a person can intuit which genre(s) the movie falls into and compare that to his or her individual preferences. While the genre itself may not be explicitly mentioned in the plot summary, many have hypothesized that a hidden representation of genre information lies within the underlying text [1]. Our goal is to confidently classify these hidden representations to their corresponding genre(s) using various NLP techniques.

A. Motivation

A genre classifier has many potential industry applications. A simple application would be for film studios to validate their work when creating plot summaries, thereby ensuring that the plot resonates with the intended genre(s). Movie streaming companies such as Netflix could leverage soft genre predictions (predicted probability of a plot summary belonging to a certain genre) as a feature within their movie recommendation system; for example, comparing the soft genre prediction of a movie to the soft genre distribution of a user's viewing history. Another application could be using patterns in soft predictions to discover new hybrid categories for genres (e.g. science fiction westerns).

II. Background

A. Related Work

There exist a number of studies that perform movie genre classification using a variety of sources including visual, audio and textual features from trailers, posters, and texts [1]. With regards to using NLP methods to classify movie genre(s) from plot summaries, there are three notable publications that we found:

[1] uses a bi-directional LSTM network to classify movie genre(s) from plot summaries for four unique genre types. They chose to first divide each plot summary into sentences and assign the genre of the corresponding movie to each sentence, thereby estimating the genre(s) for each sentence separately. They fuse the predictions for individual sentences together and use majority voting to predict the genre label for the plot. This model, however, was designed only to accommodate single-label classification, meaning only one genre can be assigned to each plot.

[2] uses several Machine Learning methods to predict movie genres based on plot summaries for 20 unique genre types. The most advanced of these Machine Learning methods is the implementation of Recurrent Neural Networks. Unlike [1], each token is evaluated at the summary-level and there is no separate sentence-level analysis.

[4] is the only paper we found that uses an attention mechanism to predict movie genres based on plot summaries. They employ self-attention to learn the importance of each word in the summary, essentially enabling the more important words to hold more weight when classifying the genre(s). Similar to [2], each token is evaluated at the plot-level and there is no separate sentence-level analysis. The scope of this study includes nine unique genre types.

Our work does not build on just a single paper. Rather we aim to extend the current progress in this field by applying recent NLP modeling innovations from other domains. Specifically, we aim to tailor Hierarchical Attention Networks (HAN) and Universal Sentence Encodings (USE) to movie genre classification.

B. Hierarchical Attention Networks

The Hierarchical Attention Network, originally proposed by [5], is designed to construct a document representation by first building representations of sentences and then aggregating those into a document representation. The HAN includes two levels of attention mechanisms — one at the word level and one at the sentence level — that let the model pay more or less attention to individual words and sentences when constructing the representation of the document [5]. For the context of multi-label genre classification, each plot summary serves as the "document" of text.

C. *Universal Sentence Encodings*

The Universal Sentence Encoder consists of models for producing sentence embeddings that demonstrate good transference to a number of other NLP tasks [7]. Experiments conducted by [7] showed models that make use of sentence level transfer learning tend to perform better than models that only use pre-trained word embeddings. The success of this sentence encoder in transfer learning tasks presents a fascinating opportunity for integration with the HAN.

In the traditional HAN, we generally use word embeddings to build sentence representations for each sentence in a document. Replacing the HAN's sentence vector construction with pre-trained sentence embeddings from the USE introduces an interesting tweak to the original architecture. The second layer for the HAN instead ingests data from the pre-trained sentence encoder. We explore whether this architecture (designated USE + HAN) can improve baseline scores in the movie genre-classification domain.

III. Data and Methods

A. *Data Description*

Our dataset is a subset of the data used in [2]. This dataset was originally sourced and parsed from raw IMDb[1] data files for movie plot summaries and their corresponding set of genres [3]. There are two data points of interest in this corpus: (1) the raw text plot summary and (2) the labeled genre(s).

The original dataset consists of 255,853 plot summaries and their corresponding genre(s). There are 20 unique genres in the original dataset. We choose to limit the scope of our analysis to ten of these genres, due to very few observations in some of the classes. After removing plot summaries that did not have at least one of the ten genre labels in our study, we are left with 224,303 unique records with an average 1.83 labels per plot. General information about the size of the summaries can be found in Table 1.

TABLE I
General Description of Plot Summaries

Average words per summary	80 words
Median words per summary	72 words
Average sentences per summary	4.4 sentences
Median sentences per summary	4 sentences

We divide the dataset into a train, validation, and test set according to an 80/10/10 split. From Figure 1, we see that the data is notably skewed, even after removing the ten least represented genres. There is over an 8:1 difference in counts between the most frequent and least frequent genre (Drama: Sci-Fi). The highly skewed data was not addressed in [2], making their models biased towards popular genres such as drama or comedy. We aim to address this issue by using more balanced subsets of the data to train a homogenous ensemble of classifiers.

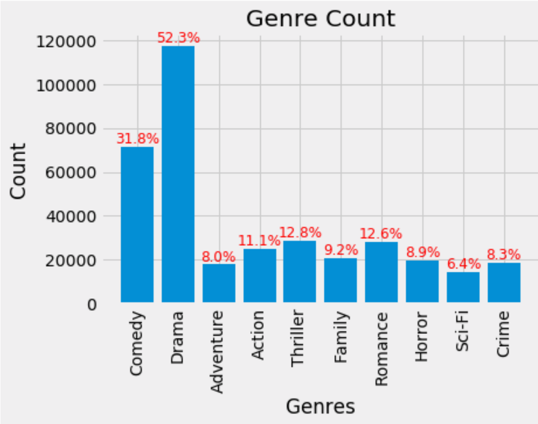

Figure 1. Full Data Genre Distribution

B. *Sampling and Homogenous Ensemble Methods*

To reduce the skew in our data, we constructed a custom multi-label sampling function based on the count **n** of the least represented genre in our training data. For each genre **g,** we subset the original training data to only include records where the label contains **g**, and take **n** samples without replacement. As the data is multi-label, our sampling technique ensured every genre had at least 10% representation but did not guarantee equal proportions. The result in Figure 2 is a much more balanced dataset with a 3:1 ratio between the most frequent and least frequent genre, a considerable drop from the 8:1 ratio observed in the original data.

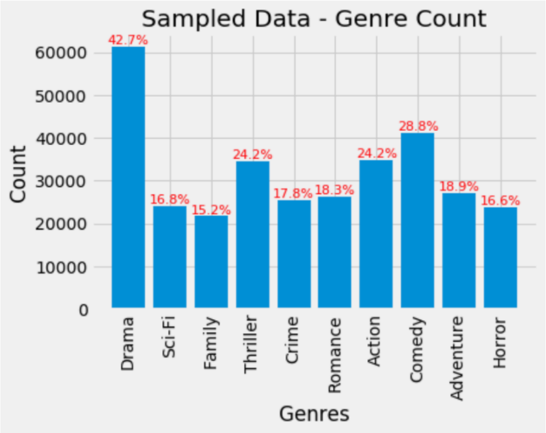

Figure 2. Sampled Data Genre Distribution

To mitigate the fact that our training set sample is considerably smaller than the entire training set, we used a homogenous ensemble approach by repeating the sampling

[1] IMDb is a prominent online database of information for films, television programs, and more

process five times and training five separate models per model-type. For model evaluation, we pass our held-out data through each of these five trained models and use majority voting to classify the genre(s) for each given plot summary. Comparable models were also trained on the full training set. A comparison of the performance between our ensemble models and the full data models is found in Section 4.

C. Pre-Processing

Before any model building took place, we performed the following pre-processing steps on our plot summaries:

1. Converted all text to lowercase
2. Eliminated all punctuation marks except ones that separate sentences
3. Converted numbers to a specific digit token
4. Converted any instance of a genre label (e.g. “Drama”) to the token “genre”
5. Lemmatized words based on their part of speech
6. Removed stop words
7. Multi-label binarized the genre labels for each plot (10-unit length binary vector where 1 indicates presence of the genre)

Steps 5 and 6 were not performed for models that included the USE, since the USE module performs best effort text input preprocessing [7]. Additionally, since the USE takes raw text strings as input instead of a tokenized list of words, we did not want to potentially disrupt its pre-trained interpretation of sentence syntax.

D. Baseline Model

Prior to constructing any models involving the HAN architecture or USE, we implemented a simple baseline model. We converted the plot summaries into a Bag-of-Words (BoW) representation and used term frequency-inverse document frequency (TFIDF) scoring to weight each of the features for a given plot summary. Even though stop words had already been filtered out, TFIDF was utilized because it could identify other terms that were common across plots and score them appropriately. The BoW representation for each plot was fed into a single-layer neural network model. The benefit of creating a BoW classifier for our experiment is to see if complex models, such as the HAN, that aim to model various intricacies of a text document will outperform a more simplistic model in multi-label movie genre classification.

E. HAN Model

Our first exploratory model was the standard HAN architecture. Our implementation used a pre-trained set of Global Vectors for Word Representation (GLoVE) to construct a vector representation for each word in each plot [6]. To normalize the plot and sentence dimensions for the HAN, we instituted a max sentence length of 15 tokens, and a max plot length of 5 sentences. Any sentences or plots that exceeded these dimensions were truncated, while any sentences or plots that fell short of these dimensions were padded. These limits were chosen to balance the competing desires of minimizing truncation of input data and minimizing the amount of padding needed. We limited our vocabulary to the 80,000 most common words in the corpus to remove all words seen two or fewer times. All out-of-vocabulary terms were converted to an “UNK” token which was mapped to the average GLoVE embedding vector.

While we defer to [5] for the comprehensive explanation regarding the components of a HAN, we provide a high-level summary below of how our plot summary vector representation is constructed. To build a sentence representation, the word embeddings for each word in the sentence are fed into a bi-directional GRU, thus creating contextually-thorough annotations for each word. Within the first attention layer, each annotation is transformed via a tanh activated fully connected layer. Word importance is measured by taking the dot product of each transformed annotation with a word level context vector and getting a normalized importance weight for each word through a softmax function. The sentence vector can then be computed as the weighted sum of the word annotations based on the normalized importance weights. This process is repeated for each sentence in a given plot. Once each sentence vector for a plot is computed via the first layer of the HAN, we compute the plot summary vector in a similar manner. Each sentence vector is passed into a bi-directional GRU to build contextually-thorough annotations. Subsequently, in the second attention layer, we transform and compute the importance of each sentence, similar to how we did in the first attention layer. After that, the plot summary vector is computed as the weighted sum of the sentence annotations based on the normalized importance weights. The plot summary vector is fed into a 10-dimensional output layer with sigmoid transformation, which outputs per-genre probabilities for multi-label classification. **Please refer to Figure A1 in the Appendix for a diagram of our HAN model architecture.**

F. USE + HAN

For our second exploratory model, we integrate the USE into the HAN architecture. We feed each sentence in the plot summary into the USE’s Deep Averaging Network (DAN). In the DAN, input embeddings for words in our sentences and bi-grams are first averaged together and then passed through a feedforward deep neural network (DNN) to produce sentence embeddings [7]. We chose the DAN encoding model because it involved less model complexity and resource consumption than the competing Transformer model [7]. Whenever the USE is mentioned in this paper, it refers to the DAN implementation.

The USE is able to accept sentences of variable length; therefore, we did not need to normalize sentence length for each sentence in a plot summary. However, we still needed to normalize the number of sentences per summary for our HAN to construct a plot summary vector. We used a max plot length of 5 sentences, similar to our first HAN model. The USE outputs a 512-dimensional vector embedding for each input string it receives, resulting in five 512-dimensional vectors per plot summary. These five sentence vectors represented an alternative to the sentence vectors output by the first layer of the traditional HAN architecture. Therefore, we removed the first layer of the HAN for this model and replaced it with the output from the USE. The second layer of the HAN that constructs the plot summary vector functions similarly to our original HAN implementation. By replacing the first layer of the HAN, we can assess whether the sentence representations generated by the USE improve the HAN's document vector construction for multi-label genre classification. **Please refer to Figure A2 in the Appendix for a diagram of our USE + HAN model architecture.**

G. USE + Fully Connected layer

Our final exploratory model (designated USE + FC) replaced the HAN architecture in its entirely by re-configuring the input to the USE. Rather than feed each individual sentence of a summary into the USE, we instead input the entire summary, resulting in a 512-dimensional representation of the entire plot. This representation was passed through a fully-connected layer before classification via a sigmoid output layer.

In addition to sentences, the USE is optimized for short paragraphs. Therefore, based on the distribution of plot summary lengths, we were not concerned with feeding entire summaries into the USE. We chose this architecture in order to make a comparison between it and the USE + HAN model to judge the impact of the HAN's sentence and document vector building layers on classification in this domain. **Please refer to Figure A3 in the Appendix for a diagram of our USE + FC model architecture.**

IV. Results and Discussion

A. Evaluation Metrics

Based on trends observed during a literature review for multi-class and multi-label genre classification, we chose to use precision, recall, and F1-score as our evaluation metrics. These were the most popular metrics used in previous studies regarding genre classification from movie plot summaries. We computed two versions of precision, recall, and F1-score: micro-average and weighted-macro-average. The micro-average metrics aggregate the contributions of all classes to compute the average metric whereas the weighted-macro-average metrics weight each class's contribution to the average by the relative number of examples available for it. These metrics were selected due to their robustness to class imbalance. In multi-label genre classification, we do not have a preference between precision and recall; therefore, our most important metric is the f1 score.

B. Experimental Setup

For each model, we conducted hyperparameter tuning to determine an optimal number of epochs for training, hidden dimensions for dense layers, dropout regularization parameter, batch size, etc. We used binary cross-entropy as our loss function and trained all Bi-GRU networks with stochastic gradient descent using an Adam optimizer. To generate genre predictions, we established a 0.5 probability threshold for each class. Since our output layer is sigmoid-transformed, the prediction for each genre is independent of the other genres, resulting in instances where our model may predict no genres for a plot summary (i.e. no genre clears the 0.5 threshold). In those instances, the genre with the highest posterior-probability is taken as the model's prediction.

C. Results

Our models' scores on the test set are reported in Table II. We purposefully exclude Weighted-Macro Recall score from the table because it is equivalent to Micro-Recall in our multi-label classification setting. The highest score for each metric and group is bolded.

For the models trained on the full training set, the precision was considerably better than the recall. We hypothesize that this is due to the skew in our data; since our models are trained on mostly dramas and comedies, they are likely to predict drama and/or comedy on the held-out data. This produces more false negatives compared to false positives, thus bringing down recall scores. Manual inspection of confusion matrices confirmed this hypothesis. The relationship between precision and recall is similar to the final results from [2].

TABLE II
Multi-Label Genre Classification Results – Test Data

Group	Model	Micro-Precision	Weighted-Macro Precision	Micro-Recall	Micro-F1	Weighted-Macro F1 Score
Ensembles with Sampled Data	BoW-TFIDF	0.576	0.574	0.619	0.597	0.595
	HAN	0.512	0.532	**0.704**	0.593	**0.603**
	USE + HAN	0.576	0.585	0.623	**0.6**	0.6
	USE + FC	**0.594**	**0.59**	0.567	0.58	0.576
Single Model – Full Training Data	BoW-TFIDF	0.695	0.676	0.487	0.573	0.542
	HAN	0.632	0.621	**0.539**	**0.582**	**0.559**
	USE + HAN	**0.705**	**0.685**	0.489	0.577	0.540
	USE + FC	0.697	0.672	0.454	0.549	0.501

Comparing the ensemble models to those trained on the full training set, we observe improvements. While precision scores drop, the recall considerably increases for all models. The increase in recall is more significant than the drop in precision, because there is an overall increase in both micro and weighted-macro F1 score across all five models. The increase in both recall and F1 score is encouraging and indicates that our models are no longer refraining from predicting the lesserrepresented genres. However, this increase in recall comes at a cost, as more false positives are predicted from the ensemble of classifiers.

Focusing on the ensemble group, there is not much difference between the baseline BoW-TFIDF model and the best performing exploratory ones. This indicates that building a plot summary vector by assuming that different words and sentences in a document are differentially informative brings about minimal improvement over a TFIDF representation.

The USE + HAN outperforms all other models in terms of micro-F1 score, while the traditional HAN outputs the best weighted-macro F1 score. While the models are quite similar from an F1-score perspective, the USE + HAN demonstrates more balance between the amount of false positives and false negatives when compared to the HAN model, which has a much higher recall compared to precision. This leads us to believe that the sentence representations produced by the USE may be superior in terms of genre representation to those produced by the bottom layer of the HAN. Additionally, since the USE + FC model performed the worst of any ensemble model on both F1 scores, we believe that the document vector constructed by the top layer of the HAN provides a significant positive impact on genre classification.

Digging into the model predictions, we discovered that the HAN predicts a greater number of genres than the other models with an average of 2.2 genre labels per plot summary, which is nearly ½ label more than the USE + HAN model and almost ¾ label more than the USE model (the average number of genres in the test set is 1.8). We hypothesize two possible explanations for this behavior. The first is that the individual HAN models that make up the ensemble are not well aligned, producing several competing genres after voting. The second is that there is some systematic feature of the HAN that tends to predict a greater number of genres. Since we do not witness this behavior in the USE + HAN model, this suggests the sentence representation as a likely cause. Further investigation into this matter could be insightful.

After noting this difference in the number of genres predicted, we also investigated the degree of agreement between the predictions of our three models by calculating the average cosine similarity of our output soft label vectors. The models were well aligned with average cosine similarities ranging from 0.89 (HAN and USE + FC) to 0.96 (USE + HAN and USE + FC). This was somewhat surprising considering the high dimensionality of the output space and the moderate precision and recall values.

Since the performance of the ensemble models is relatively close, it is also interesting to note the variation in training time. The HAN ensemble took the longest, requiring about 30 minutes of preprocessing and about 2.5 hours for ensemble training, while the USE + FC was the fastest, requiring about 30 minutes to load the sentence encodings and less than 5 minutes to train the ensemble.

D. Discussion of Errors

Error analysis helps us determine assumptions and learnings of the models that led to erroneous classifications. This knowledge could be used in future work to further advance this body of research. Table III contains multiple examples of plot summaries and the predictions output (w/ average level of certainty) by each of our three exploratory ensembles.

Examining example 1, we see an instance of ambiguity. The user who annotated this movie's genre chose "Horror" and "Thriller." These two genres are often linked and not clearly delineated in many movie-goers' minds. Based on the plot summary, we would have classified this film as just "Horror", a classification shared by the HAN and USE + HAN models. Since genres do not have clear definitions and the genre labels were given by a wide variety of users each of whom had his or her own interpretation, the data set contains many instances like this where subjectivity plays a large role.

TABLE III
Multi-Label Prediction Examples – Exploratory Models

Plot Summary	Predicted Genre(s)	Actual Genre(s)
(1) Five youngsters are on their way back home. They stop the car for no reason and they begin to disappear, one by one. Some of them leave a trail of blood behind. There's a sick maniac around who just wants torture them, killing them and have a little fun with their misery. The camera they bring with them records everything. Until the end...	HAN Horror (0.98) USE + HAN Horror (0.88) USE + FC Horror (0.84) Thriller (0.63)	Horror Thriller
(2) A couple and a dog. They live not far from Mexico City. Strange things happen in the house. She is in love. She'd like to be pregnant. He is busy and tired. The woman is undressing inside the house. He falls asleep. She pours wine on his face. A strange visitor dances outside. She stares at us, under the waterfall. They are imagining things. They might be happy in a near future. The film was inspired by a famous Edward Hopper painting, 'Summer evening'.	HAN Drama (0.51) USE + HAN Drama (0.53) USE + FC Comedy (0.13)	Drama
(3) After 30 years together, Henry's parents are finally getting married. As the responsible oldest son he wants everything to be perfect on their big day. But his youngest brother Tom can barely stay conscious and George (the middle brother) can't stop thinking about his ex-wife long enough to focus on the wedding. To complicate things further, Rhonda, the wedding coordinator chooses today to tell George that she's loved him for 10 years. Can Henry pull everyone together in time for the wedding? And what will his parents do if he can't?	HAN Comedy (0.95) Romance (0.67) USE + HAN Comedy (0.95) Romance (0.67) Drama (0.53) USE + FC Comedy (0.51) Drama (0.62)	Comedy

Additionally, since many different users created the plot summaries, the style, quality, and accuracy of each summary varies wildly. Example 2 showcases a poorly written summary. A human reader likely will classify this as a Drama but will not be quite sure as the summary has several elements (strange events at the house, strange person dancing) that are not easily comprehended. In this case, two of the models correctly classified it though neither were very confident. Having a more consistent dataset would be a huge boon to furthering genre classification.

Example 3 shows the potential negative role of attention. The HAN and USE + HAN models incorrectly label this movie as a "Romance" likely due to the presence of words related to marriage as these words are usually seen in "Romance" movies. The USE + FC model, on the other hand, lacks any attention mechanism and, while not correctly predicting only "Comedy", does not make the mistake of classifying this as a "Romance."

V. Conclusion

Achieving a significant improvement over a BOW-TFIDF model proves to be challenging in the multi-label movie genre classification domain. While usage of the Hierarchical Attention Network architecture and/or Universal Sentence Encodings do achieve improvements in some of our target metrics, no model we explored was a clear winner across the board. However, from a training efficiency standpoint, the USE + FC architecture was significantly faster. It may be worthwhile to investigate other possible architectures that start with Universal Sentence Encodings for this domain.

Additionally, future work in this domain would strongly benefit from an investment in quality training data. Many of our errors could be at least partially attributed to subjective labeling of genres and/or poorly written plot summaries. Finally, utilizing sampling to achieve a more balanced training set led to a considerable improvement in F1 score. However, even sampling did not completely eliminate class imbalance. A more balanced training set would be helpful in reducing the risk that the model is just predicting the most common labels.

VI. References

[1] Ertugrul, Ali Mert & KARAGOZ, Pinar. (2018). Movie Genre Classification from Plot Summaries Using Bidirectional LSTM. 10.1109/ICSC.2018.00043.

[2] Hoang, Q. (2018). Predicting Movie Genres Based on Plot Summaries. *CoRR, abs/1801.04813*.

[3] Imdb data. ftp://ftp.fu-berlin.de/pub/misc/movies/database/frozendata/

[4] Wehrmann, Jonatas, Lopes, Mauricio, AND Barros, Rodrigo. "Self-Attention for Synopsis-Based Multi-Label Movie Genre Classification" *Florida Artificial Intelligence Research Society Conference* (2018)

[5] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation

[7] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil. Universal sentence encoder. CoRR, abs/1803.11175, 2018. URL http://arxiv.org/abs/1803.11175.

[8] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.

[9] A.(2017, October 5) anargyri/lstm_han. Retrieved from https://github.com/anargyri/lstm_han

[10] Hoang, Q. (2018) qhoang/Movie-Genres. Retrieved from https://github.com/qhoangdl/Movie-Genres

VII. Appendix

In this section we provide diagrams for the three exploratory models utilized in this paper

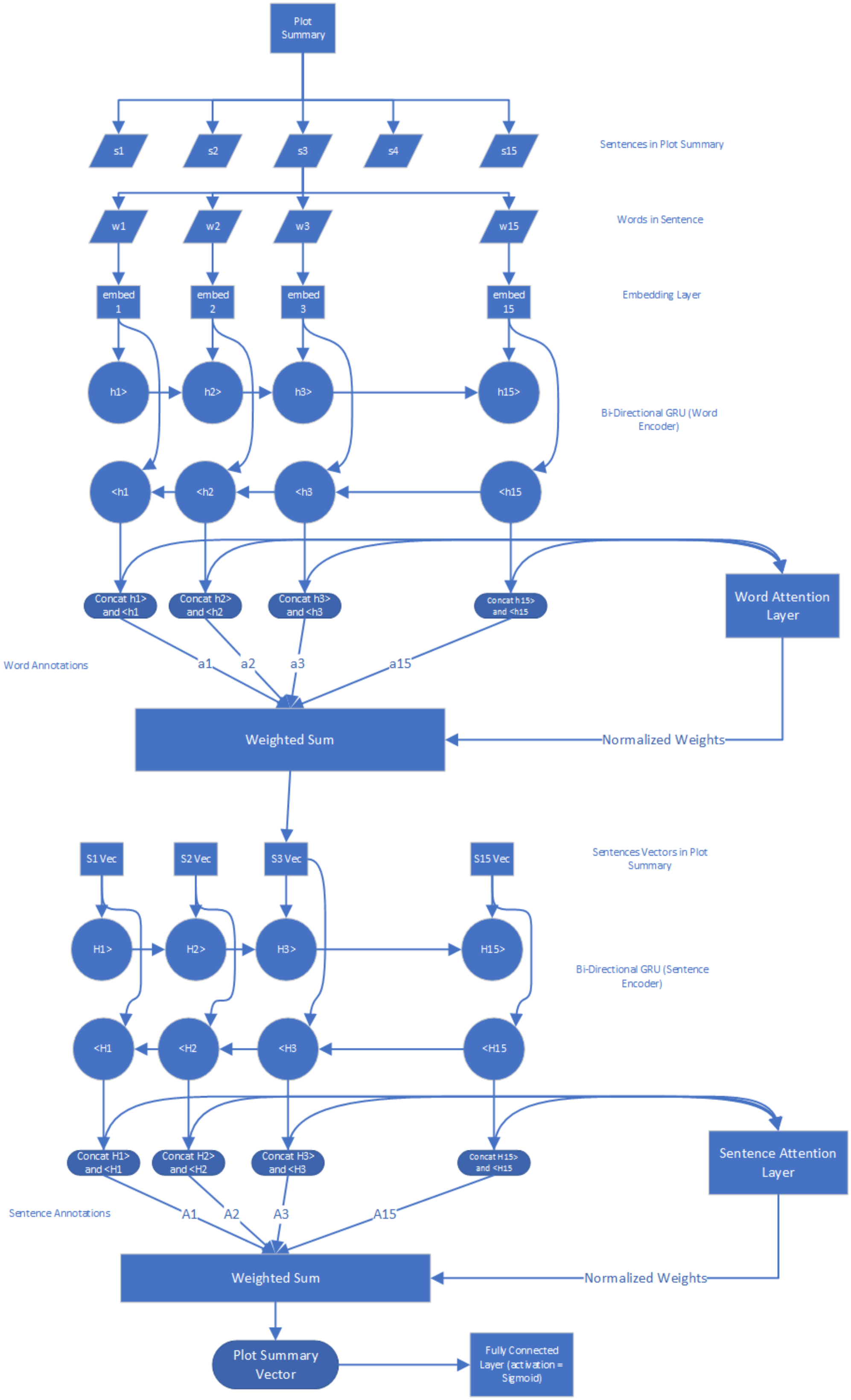

Figure A1. HAN architecture

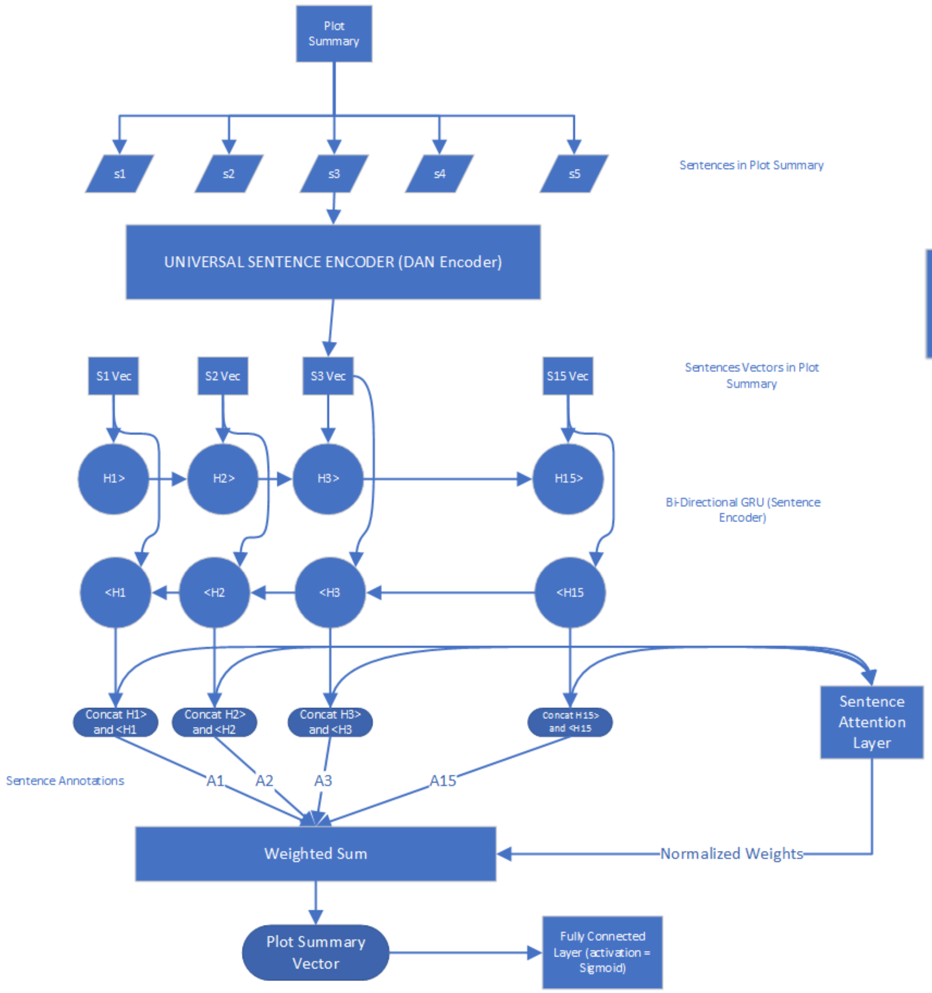

Figure A2. USE + HAN architecture

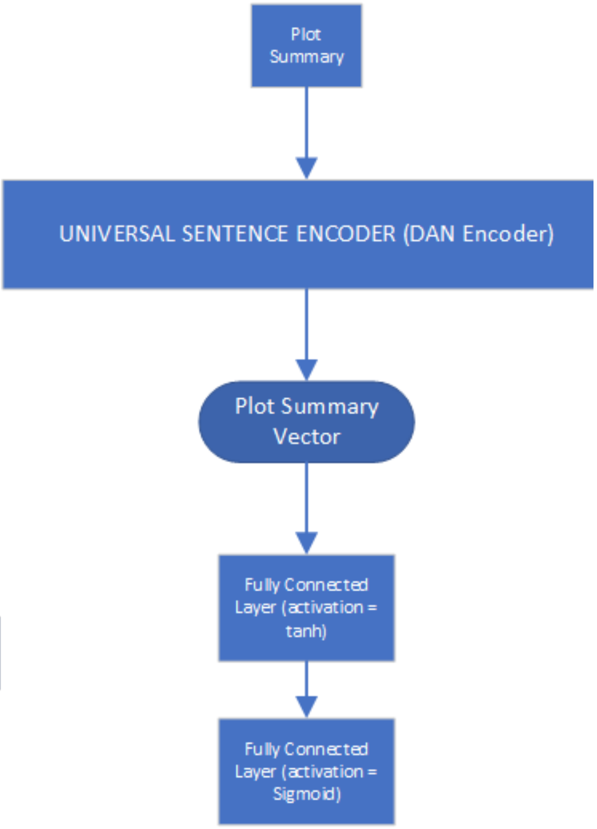
Plot Summary
UNIVERSAL SENTENCE ENCODER (DAN Encoder)
Plot Summary Vector
Fully Connected Layer (activation = tanh)
Fully Connected Layer (activation = Sigmoid)

Figure A3. USE + FC architecture