Figuratively Speaking: Authorship Attribution via Multi-Task Figurative Language Modeling

Gregorios A Katsios[*] and **Ning Sa**[**] and **Tomek Strzalkowski**[*,**]
*Department of Computer Science, **Department of Cognitive Science
Rensselaer Polytechnic Institute
{katsig, san2, tomek}@rpi.edu

Abstract

The identification of Figurative Language (FL) features in text is crucial for various Natural Language Processing (NLP) tasks, where understanding of the author's intended meaning and its nuances is key for successful communication. At the same time, the use of a specific blend of various FL forms most accurately reflects a writer's style, rather than the use of any single construct, such as just metaphors or irony. Thus, we postulate that FL features could play an important role in Authorship Attribution (AA) tasks. We believe that our is the first computational study of AA based on FL use. Accordingly, we propose a Multi-task Figurative Language Model (MFLM) that learns to detect multiple FL features in text at once. We demonstrate, through detailed evaluation across multiple test sets, that the our model tends to perform equally or outperform specialized binary models in FL detection. Subsequently, we evaluate the predictive capability of joint FL features towards the AA task on three datasets, observing improved AA performance through the integration of MFLM embeddings.

1 Introduction

Figurative Language (FL) constructs, such as metaphor, simile, and irony, are common in various forms of communication, such as literature, poetry, and speech. Their use can enrich the meaning, creativity, and persuasiveness of a message and help to achieve an intended impact on the reader. The use of certain forms of FL in writing reflects the authors' style and background, including their education, personality, social context, and worldviews. Therefore, we hypothesise that the choice of figurative language features in (written) communication may reveal the writer's cognitive and linguistic basis that underlie their production, and how their selection is influenced by the context, the intention, and the emotion of the writer.

In this paper, we introduce a multi-task classification model designed to detect multiple Figurative Language (FL) features in a body of text. The first research question (RQ1) we seek to answer is: "Is a model that is trained to detect multiple FL features simultaneously more effective than multiple specialized models, each trained to detect a specific FL feature?" Through our research, we demonstrate that this multi-task model is indeed more effective than using several binary models.

In our research, we utilize 13 publicly available datasets to train and evaluate both binary and multi-task models. We deliberately opted against integrating additional datasets specifically designed for metaphor detection, which is only one of the phenomena we study. The rationale behind this decision was creating a more balanced training data, which otherwise would have been disproportionately skewed our study towards metaphor detection, given the substantially more resources dedicated to this phenomenon. At the same time, the lack of annotated corpora for other figurative language features such as personification, metonymy, oxymoron, etc. necessarily limited our initial study to the six FL constructs that are generally well represented among these 13 datasets: Metaphor, Simile, Idiom, Sarcasm, Hyperbole, and Irony.

All our binary models and the multi-task model are based on RoBERTa (Liu et al., 2019). After training the specialized binary models on the combined datasets, we used them to automatically label our training corpora with all applicable FL features. This multi-label dataset was then used to train our multi-task model. Afterwards, we compare our Multi-task Figurative Language Model (MFLM) against the binary classifiers on the 13 test sets. The results showed that MFLM matched or outperformed the binary classifiers in five test sets and achieved higher task-specific performance than the binary models in another three test sets, which suggests that these features are not independent from one another.

Findings of the Association for Computational Linguistics ACL 2024, pages 13240–13255

After training our multi-task figurative language classifier, we put forward a second research question (RQ2): "Does the incorporation of Figurative Language (FL) features enhance performance in Authorship Attribution (AA) tasks?" To answer this, we evaluate the impact of the FL features learned by our Multi-task Figurative Language Model (MFLM) on three publicly available AA datasets, each consisting of documents with varying topical content and number of authors. For each dataset, we train Multi-Layer Perceptron (MLP) classifiers, using MFLM sentence embeddings and other baselines as input features. The baselines consist of classical Stylometric features, character and word n-gram TF-IDF vectors, and generic sentence embeddings. Our results demonstrate that the AA task performance is indeed improved by combining MFLM embeddings with other baselines.

To our knowledge, this work is the first to examine the applicability of FL features in AA. We should note here that we did not expect that the FL features alone would be sufficient to perform AA; rather we set off to demonstrate that incorporating combined FL features improves AA performance when integrated with more basic stylistic features, particularity for longer texts. The results show that the latter is generally true; however, we found that the FL features perform nearly as strong and sometimes better on their own. This supports our initial stipulation that FL use is highly personalized, and thus an excellent predictor of authorship.

We make our code and data available in our GitHub repository[1].

2 Related Work

Most of the previous studies on Figurative Language (FL) feature detection focus on the features independently. An earlier work, (Tsvetkov et al., 2014), used lexical semantic features of the words to discriminate metaphors from literals. More recently, Choi et al. (2021) utilized metaphor identification theories using RoBERTa to predict whether a word in a sentence is metaphorical or not. A similar shift from linguistic feature based approach to pre-trained language model (PLM) based approach is observed in simile detection. Niculae and Danescu-Niculescu-Mizil (2014) extracted features such as topic-vehicle similarity and imageability to separate similes from literal comparisons. Ma et al. (2023) used BERT (Devlin et al., 2018) and RoBERTa in simile property probing tasks and concluded that the PLMs still underperformed humans. PLMs are also applied to the detection of sarcasm (Yuan et al., 2022), hyperbole (Biddle et al., 2021), irony (González et al., 2020), and idiom (Briskilal and Subalalitha, 2022).

Among the studies that work on more than one features, Badathala et al. (2023) used datasets cross-labeled with metaphor and hyperbole, and found that the multi-task learning approach performed better than the single-task approach on both features. Chakrabarty et al. (2022b) rendered the FL detection into a multi-task natural language inference (NLI) problem, developed a NLI dataset of four FL features, and tested with several experimental systems. Chakrabarty et al. (2022a) collected datasets on idiom and simile and developed knowledge enhanced RoBERTa-based models. However, their task was to predict the correct continuation of the given narrative, not FL feature detection. Adewumi et al. (2021) built a dataset covering 9 FL features plus literals. They tested three baseline systems in a multi-class classification task and BERT outperformed the other two systems.

There is a rich literature in the field of Authorship Attribution (AA). Various methods have been applied to the task, ranging from SVM based approaches, such as (Kestemont et al., 2018), to transformer based models, like (Bauersfeld et al., 2023). In PAN-2019 cross-domain AA challenge (Kestemont et al., 2019), most of the submissions used n-gram features (char, word, part-of-speech) and an ensemble of classifiers (SVM, Logistic Regression, etc). Fabien et al. (2020) fine-tuned a BERT model for AA task and tested the model on three datasets including IMDB-62 (Seroussi et al., 2014). In a recent review article (Tyo et al., 2022), feature based methods and embedding based methods were tested and compared on the same datasets. They used n-grams, summary statistics and co-occurance graphs as features, as well as static char/word embeddings and transformer-based sentence embeddings.

3 Figurative Language Modeling

In our study, we investigate the potential benefits of combining Figurative Language (FL) features as opposed to analyzing each feature independently. To answer our first research question, we examine whether training a FL classification model capable of jointly labeling text with relevant features would

[1]Figuratively Speaking: https://github.com/HiyaToki/Figuratively-Speaking

outperform a singular binary model specialized in detecting only one feature. This idea stems from noticing that in both spoken and written language, individuals intertwine various elements of figurative speech to effectively convey their intended message. Consequently, FL features frequently co-occur, and understanding the interplay between these features may offer valuable insights for improving their identification accuracy. This research builds upon prior studies that explored the simultaneous detection of metaphors and sarcasm, as well as hyperbole and sarcasm. In our investigation, we aim to simultaneously learn to detect six distinct FL features: Metaphors, Simile, Sarcasm, Hyperbole, Idiom, and Irony.

3.1 Data

In our research to learn to classify FL phenomena, we rely on publicly available datasets. In total, we work with 13 individual corpora, which are summarized in Table 1 (see Appendix A for additional details). While space constraints prevent exhaustive descriptions, we encourage interested readers to explore the original works by the dataset creators for comprehensive insights into the data collection and annotation processes.

Among the datasets we analyze, the iSarcasm corpus (Farha et al., 2022) stands out as truly multi-labeled. It includes training and testing examples annotated with labels such as sarcasm, irony, overstatement (hyperbole), understatement, satire, and rhetorical questions. For instance, an excerpt from the iSarcasm training set reads: *"Can't wait to be back at uni so I can order more shoes and clothes without my mum telling me off"*, which is labeled with both sarcasm and hyperbole.

In contrast, several other datasets adopt a multi-class approach. Each example in these datasets corresponds to a single applicable label. Additionally, some datasets focus exclusively on specific FL phenomena, employing positive and negative examples (e.g., *feature_X* and *not_feature_X*) to create a binary distinction.

When dealing with FL datasets, it's crucial to consider how negative examples are constructed. Some datasets construct the negative class (i.e., *not_feature_X*) by ensuring that samples represent true literal sentences devoid of any FL speech. The FLUTE corpus (Chakrabarty et al., 2022b) is an example of this approach, where FL sentences are paired with their rephrased literal counterparts. For instance, the figurative sentence (metaphor): *"A break up can leave you with a broken heart"* is paired with the literal sentence: *"It's hurtful when a breakup makes you feel lonely and sad"*.

Other datasets annotate the negative class as simply not containing the FL phenomena described by the positive class. For instance, in the Irony SemEval 2018 corpus (Van Hee et al., 2018), sentences that are labeled as *not_irony* may still exhibit other FL traits. Consider the sentence: *"Look for the girl with the broken smile"* which, although not ironic, contains a metaphor that is not explicitly annotated.

In our pipeline, we apply minimal preprocessing to the sentences from these corpora, and we load them into our combined collection, retaining human annotations relevant to our work. Notably, we focus on the six FL features listed in Table 1, ignoring classes beyond this scope. At this stage, we clearly distinguish between literal sentences and negative class sentences labeled as *not_feature_X*.

In our study, we encounter various datasets with distinct characteristics regarding their train/dev/test splits. Some datasets come with a predefined splits, where we merge the training and development sets into a single training set, reserving the original test set solely for evaluation. In cases where datasets lack existing splits, we adopt a systematic approach, setting aside a 10% stratified sample for testing. The entire collection consists of 69168 training and 9729 testing examples.

3.2 Binary Models

To detect the various FL phenomena, we create task-specific binary classifiers. This process involves combining datasets annotated with examples relevant to each specific feature. For instance, to train a classifier for metaphors, we aggregate data from PIE-English, FLUTE, LCC, and MOH datasets. Similarly, for simile classification, we gather data from PIE-English, FLUTE, MSD23, and Figurative Comparisons datasets.

The combination of datasets allows us to establish both positive and negative sets for each classification task. In the context of training a metaphor classifier, the positive set comprises examples exhibiting metaphoric expressions, while the negative set encompasses instances without metaphors. As detailed in Section 3.1, certain datasets exclusively utilize literal examples for constructing the negative class, whereas others use examples not containing the FL phenomena described by the positive

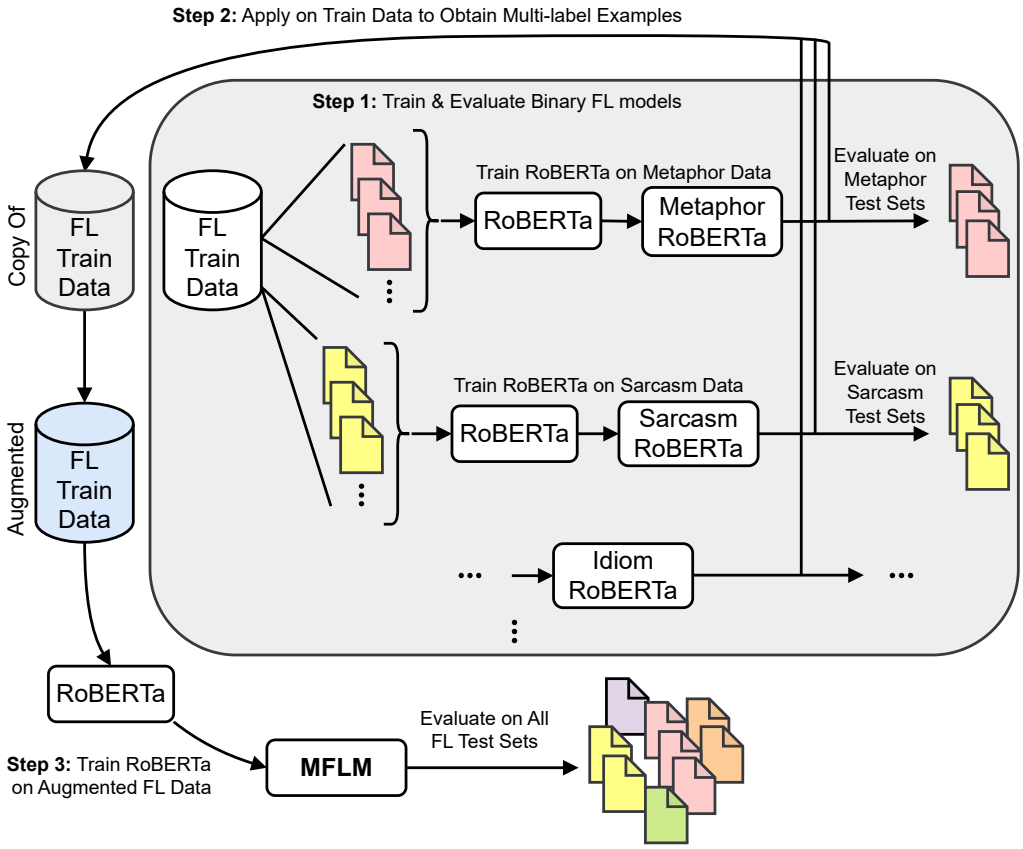

Figure 1: Diagram illustrating our pipeline of training the individual binary FL models, augmenting the FL training collection with predicted labels and fine-tuning the MFLM.

class. Thus, achieving a balanced representation necessitates the inclusion of negative samples from both types of datasets.

In our approach, positive and negative examples are retrieved from the combination datasets corresponding to the specific task, while literal examples are sourced from across all datasets. The final training set for each task is formed by selecting all positive examples and supplementing them with an equal number of negative and literal examples. Specifically, if the size of the positive class is denoted as N, we sample $N/2$ negative and $N/2$ literal examples. In scenarios where there are insufficient negative examples, we augment the dataset with an appropriate number of literal examples to ensure a total of $2N$ training instances. During training, the labels of literal examples are transformed to *not_feature_X*, aligning with our objective to create robust binary classifiers capable of discerning sentences containing the specific feature from those that do not. For detailed information on the number of training samples per task, please refer to the Appendix A.

Subsequently, we train individual RoBERTa (Liu et al., 2019) models[2] for each task using a standardized set of hyper-parameters across all training jobs: Epochs: 5, Learning Rate: 2e-5, Weight Decay: 0.01, Warm-up Ratio: 0.1, Batch Size: 16. The time required to train a binary model averaged at approximately 80 minutes, using a single NVidia RTX A6000 GPU.

3.3 Multi-Task Model

We proceed to train a Multi-task Figurative Language Model (MFLM) that can label a sentence with all applicable features in a single pass. For this, we convert our combined training dataset into a multi-label format. We use the array of binary models to assign all the possible labels to each training sentence in our corpora.

Consequently, we obtain an augmented FL training corpus, for which every sentence has a corresponding list of predicted FL labels. To produce a high quality training set, we keep only the examples where the predicted labels are consistent with the original human annotations. For

[2]RoBERTa-Large: https://huggingface.co/FacebookAI/roberta-large

Datasets	Metaphor	Simile	Sarcasm	Hyperbole	Idiom	Irony	Literal
Reddit Irony Corpus (Wallace et al., 2014)	-	-	-	-	-	537	No
Irony SemEval18 (Van Hee et al., 2018)	-	-	-	-	-	2212	No
iSarcasm (Farha et al., 2022)	-	-	846	46	-	174	No
Sarcasm Corpus (Oraby et al., 2017)	-	-	4693	1164	-	-	No
MOVER (Zhang and Wan, 2021)	-	-	-	1007	-	-	Yes
HypoGen (Tian et al., 2021)	-	-	-	1876	-	-	Yes
EPIE (Saxena and Paul, 2020)	-	-	-	-	2761	-	Yes
PIE-English (Adewumi et al., 2021)	12590	1072	46	-	13738	30	Yes
FLUTE (Chakrabarty et al., 2022b)	749	750	2677	-	1009	-	Yes
MSD23 (Ma et al., 2023)	-	3576	-	-	-	-	Yes
Figurative Comparisons (Niculae and Danescu-Niculescu-Mizil, 2014)	-	449	-	-	-	-	Yes
LCC (Mohler et al., 2016)	3036	-	-	-	-	-	No
MOH (Mohammad et al., 2016)	410	-	-	-	-	-	Yes

Table 1: Datasets used in our multi-task Figurative Language approach. The values in the cells denote the number of examples per feature. The last column indicates whether a dataset employs literal examples to form the negative class. Blank fields correspond to datasets (rows) that do not contain any annotations for the corresponding FL feature (columns).

instance, if a sentence is annotated by humans as: *[metaphor, idiom]*, we accept predictions such as: *[metaphor, idiom, simile, not_irony, not_hyperbole, not_sarcasm]*, but we reject predictions like: *[metaphor, not_idiom, not_simile, not_irony, not_hyperbole, not_sarcasm]*, due to the *not_idiom* prediction's inconsistency.

In this manner we create a dataset of 61264 sentences, discarding 7904 text-prediction pairs that conflict with human annotations. The distribution of labels in the dataset is shown in Table 2. We allocate 10% of this training set to be used as a development set, facilitating the identification of the optimal probability threshold for each feature. Leveraging both automatically generated labels and human annotations, we obtain two distinct sets of thresholds. One set is optimized based on the human labels, while the other set is calibrated using the automatic labels.

Metaphor	18981
Simile	6618
Sarcasm	9906
Hyperbole	13699
Idiom	18604
Irony	10176

Table 2: Class distribution of the automatically annotated multi-label training dataset.

We follow the same hyper-parameter set-up as the binary model training, and the average time to train the multi-task model is about 326 minutes, using a single NVidia RTX A6000 GPU. Our pipeline of training the individual binary FL models, augmenting the FL training collection with predicted labels and fine-tuning the MFLM, is illustrated in Figure 1.

3.4 Evaluation and Results

To evaluate both binary and multi-task approaches, we use the reserved task-specific testing sets. In Tables 3a and 3b, we report the weighted average F1-score obtained from a single run. The rows marked as Metaphor, Simile, Sarcasm, Hyperbole, Idiom and Irony refer to binary models while the rows marked as MFLM refer to our multi-task model. MFLM-h and MFLM-b refer to predictions acquired by tuning the probability thresholds on the development set using human annotations and binary predictions respectively.

Due to space limitations, we present a single column for the multi-class test sets. Nonetheless, our binary models were evaluated appropriately, by treating annotations from unrelated tasks as *not_feature_X*. For instance, when evaluating the Metaphor binary model on the FLUTE test set, simile, sarcasm and idiom ground truth labels become *not_metaphor*. In contrast, since our MFLM can inherently support all classes, we report weighted F1-score without altering the ground truth labels.

The MFLM demonstrates competitive or superior performance compared to binary classifiers across different test sets. Specifically, in 5 out of 13 tests, the MFLM either matches or surpasses binary models. Furthermore, in 3 tests, the MFLM exhibits comparable or superior performance in specific tasks. For instance, MFLM-h performs equally well as the Simile and Sarcasm models on the FLUTE test sets, achieving F1-scores of 0.98 and 0.97 respectively. Moreover, the MFLM-h surpasses the Sarcasm model on the Sarcasm

	FLUTE	iSarcasm	Sarcasm Corpus	MSD23	Figurative Comparisons	LCC
Metaphor	0.76	-	-	-	-	0.83
Simile	0.98	-	-	0.80	0.81	-
Sarcasm	0.97	0.82	0.80	-	-	-
Hyperbole	-	0.96	0.56	-	-	-
Idiom	0.85	-	-	-	-	-
Irony	-	0.79	-	-	-	-
MFLM-h	0.87	0.43	0.70	**0.84**	0.78	0.67
MFLM-b	0.89	0.37	0.72	**0.83**	**0.81**	0.74

(a) Part 1 of the evaluation results.

	MOH	EPIE	PIE-English	Irony SemEval18	Reddit Irony	HypoGen	MOVER
Metaphor	0.81	-	0.92	-	-	-	-
Simile	-	-	0.98	-	-	-	-
Sarcasm	-	-	-	-	-	-	-
Hyperbole	-	-	0.86	-	-	0.70	0.71
Idiom	-	0.91	0.99	-	-	-	-
Irony	-	-	0.97	0.67	0.66	-	-
MFLM-h	0.58	**0.91**	0.97	**0.76**	0.49	0.69	0.64
MFLM-b	0.58	**0.94**	0.97	**0.71**	0.52	**0.77**	0.64

(b) Part 2 of the evaluation results.

Table 3: Evaluation results on task-specific test sets. We report the weighted F1-score. With bold we draw attention to evaluation results where the MFLM is on par or surpasses the corresponding binary model. Scores that are underlined correspond to cases where the MFLM is on par or outperforms a binary model on a specific task. Blank fields correspond to binary models (rows) that are not applicable to the corresponding test set (columns).

Corpus test set, with F1-scores of 0.82 and 0.80 respectively. On the same test set, the Hyperbole model outperforms the MFLM-h in the hyperbole task, with F1-scores of 0.56 and 0.33 respectively. In the PIE-English test set, the MFLM-h excels over the Metaphor binary model on the metaphor task with 0.96 versus 0.92 F1-score respectively, and matches the performance of the Idiom model. This supports our first research question and highlights the versatility and effectiveness of the MFLM across different linguistic tasks and datasets.

3.4.1 Error Analysis

To pinpoint the weaknesses and strengths of our MFLM, we conduct a manual error analysis, scrutinizing samples where the multi-task and/or binary models disagree with the ground truth. For each case, we display a few random examples in Table 4, while more samples are presented in the Appendix A for further reference. Our findings indicate that the majority of miss-classifications made by the MFLM stem from inaccuracies or incompleteness in the annotation of input sentences. Nonetheless, the predictions generated by the MFLM demonstrate a reasonable level of accuracy in most instances and carry on to experiment using our proposed multi-task FL model and evaluate its appropriateness on the Authorship Attribution (AA) task.

4 Authorship Attribution

We proceed to investigate the effectiveness of our Multi-task Figurative Language Model (MFLM) in the closed-case Authorship Attribution (AA) downstream task. AA involves classifying texts to determine their respective authors from a known set of candidates. Specifically, given a training corpus consisting of N authors, the objective is to predict the author of each document in the test set by selecting from the set of N authors.

Our second research question proposes that embeddings incorporating figurative language features will enhance performance in the AA task. This concept extends from stylometric analysis (Lagutina et al., 2019), which traditionally concentrates on discerning patterns within written text. Stylometric analysis examines various aspects of writing style, including word selection, sentence construction, punctuation usage, and vocabulary preferences. To the best of our knowledge, our study is the first of its kind to utilize a Transformer model that has been fine-tuned for multi-task FL classification, towards the AA task. Previous research in this area minimally explored the applicability of FL features for this specific task.

4.1 Data

In our Authorship Attribution (AA) experiments, we employ three distinct, publicly acces-

MFLM & Binary models disagree with GT

GT	Literal	*The guests **showered rice** on the couple.*
MFLM	Metaphor	
Bin	Metaphor	
GT	Literal	*Charlie'd asked me if I'd like to **make a bit on the side**.*
MFLM	Metaphor, Idiom	
Bin	Metaphor	
GT	Not Irony	*And **seeing the light on the current drug policies**.*
MFLM	Metaphor, Idiom, Irony	
Bin	Metaphor, Idiom, Irony	

MFLM disagrees with GT, Binary models agree with GT

GT	Literal	*She said he was very nice and he **beamed a smile** at her.*
MFLM	Metaphor	
GT	Literal	*I was talking to someone and **we had great chemistry** then they **went ghost on me**.*
MFLM	Idiom	
GT	Literal	*I can't get pregnant but all my friends are having kids.*
MFLM	Irony	

Table 4: Error analysis - samples where model predictions do not align with human annotations (ground truth).

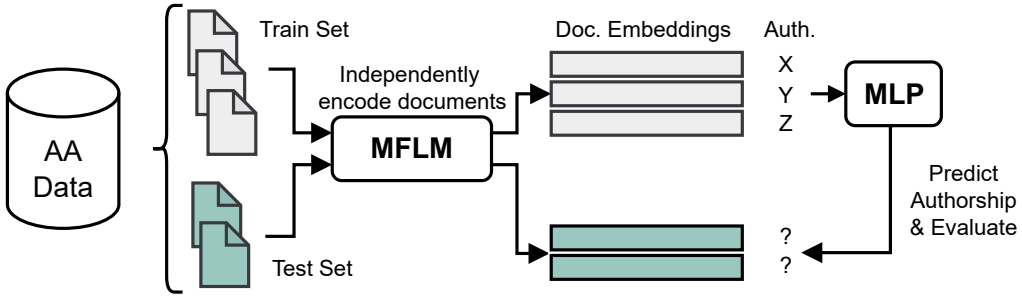

Figure 2: Diagrammatic representation of our Authorship Attribution training and evaluation approach. Following this process, any baseline can take the place of the "MFLM" rectangle.

sible datasets. The first dataset, IMDb-62 (Seroussi et al., 2014), comprises 1000 movie reviews from each of the 62 authors. These reviews are relatively short, averaging around 100 words. The IMDB-62 dataset does not have a predetermined train/test split, therefore we reserve a 10% stratified sample for testing. This yields a training set of 55800 examples and a testing set of 6200 samples.

The second dataset, PAN-2006 (Houvardas and Stamatatos, 2006), is focused on corporate and industrial topics. It includes short texts of approximately 500 words. The training set comprises 2500 texts, with 50 texts per author. Similarly, the test set consists of 2500 texts, with 50 texts per author, ensuring no overlap with the training data.

The third and final dataset, PAN-2018 (Kestemont et al., 2018), contains medium-length texts of around 800 words each, centered on fan fiction. This dataset is divided into four problems, each with a different number of authors (20, 15, 10, and 5). However, each author consistently contributes seven texts. The test sets vary in the number of texts they contain, with 79, 74, 40, and 16 texts respectively. In our experiments, we use only the English texts.

4.2 Baselines

In our Authorship Attribution (AA) task, we evaluate the performance of our MFLM against four different baselines. The first baseline is built upon classical Stylometric features. We implement 52 text metrics using the cophi[3] and textstat[4] Python packages. These metrics are used to form a document vector with 52 stylometric features. For a more detailed explanation of these features, please refer to the Appendix A.

The second baseline utilizes the all-roberta-large-v1[5] Sentence Embedding (Reimers and Gurevych, 2019) model, which we refer to as SBERT in the following sections. This model is comparable to our MFLM since it is also based on RoBERTa-Large, but without the multi-task FL classification

[3]cophi: https://github.com/cophi-wue/cophi-toolbox

[4]textstat: https://github.com/textstat/textstat

[5]all-roberta-large-v1: https://huggingface.co/sentence-transformers/all-roberta-large-v1

fine-tuning. With SBERT, we generate a 1024-dimensional document vector. This vector is computed by averaging the individual sentence embeddings for each input text.

The third and fourth baselines in our study are constructed using word and character n-grams, respectively. We utilize the Python package scikit-learn (Pedregosa et al., 2011) to analyze the texts and identify the 1024 most common n-grams from the training dataset, where the value of n varies from 1 to 5. We exclude stop words from the input texts during this process. Subsequently, we compute the Term Frequency-Inverse Document Frequency (TF-IDF) values for these n-grams across all documents, resulting in 1024-dimensional sparse document vectors.

4.3 Evaluation and Results

For the evaluation, we begin by encoding all texts in the AA datasets utilizing our MFLM model and the baselines. To create the embeddings using the MFLM, we discard the multi-task classification layer and directly utilize the underlying Transformer model. The sentence embedding is computed by mean-pooling all token embeddings, including the *[CLS]* token, taken from the last hidden layer. To create the document embedding, we average the embeddings of individual sentences. This allows us to create a 768-dimensional vector for each document.

Following this encoding step, we construct Multi-Layer Perceptron (MLP) classifiers for each test case and features combination. These MLP models consist of a single hidden layer comprising 1024 units and are implemented using the Python package scikit-learn. Our training process involves 1000 epochs with a learning rate of 2e-5, incorporating early stopping. The activation function employed is ReLU (Agarap, 2018), and the optimizer used is Adam (Kingma and Ba, 2014). Subsequently, we apply the trained model on the test set to calculate weighted average F1-scores obtained from a single run, which are presented in Table 5. The training and evaluation process for Authorship Attribution (AA) is illustrated in Figure 2.

Character and word n-grams features remain a valuable tool for AA, as their strength lies in capturing stylistic features like word choice, punctuation, and common phrases, often unique to an author. N-gram features, encompassing character sequences, spelling preferences, and even made-up words, remain consistent even with smaller datasets and paraphrasing. This robustness makes them effective for identifying rare words, misspellings, and author-specific quirks. However, they lack the ability to capture the semantic and pragmatic aspects of meaning or structural organization of text (which we do not address in this paper), both essential aspects of an author's overall style.

On the other hand, MFLM document vectors address both semantic and pragmatic aspects by encoding Figurative Language (FL) features within sentences. This approach allows for a more nuanced comparison of texts, considering not only the use of metaphors, similes, and other rhetorical devices by the author, but also their unique combinations. This could potentially lead to a more effective generalization across various writing styles and genres. Prior work on FL and metaphors (Lakoff and Johnson, 2008; Thibodeau et al., 2009) has noted that authors often blend their FL constructs in a seemingly haphazard manner. Rather than conforming to any discernible "logic", this pattern seems to be a reflection of the author's individual style, as suggested by our findings. While quite powerful, FL-based features don't encompass all facets of an individual's writing style. We continue to investigate the structural aspects of texts, which is one area that remains under study. On the other end of the spectrum, we must also account for information contained in subword patterns, an area where n-grams excel. Additionally, typos, grammatical errors, and paraphrasing can significantly impact MFLM embeddings, potentially resulting to misleading attributions.

Furthermore, we conducted experiments to explore the impact of integrating Figurative Language (FL) features by combining our MFLM encoding with baseline document vectors and subsequently training new MLP classifiers. Our findings demonstrate a consistent boost in performance across nearly all cases when using the combined features, thereby supporting our second research question.

In Table 5, we also include state-of-the-art (SOTA) results, as reported in (Tyo et al., 2022) and (Kestemont et al., 2018). The methodologies vary across implementations, but character n-grams, part-of-speech n-grams, and summary statistics typically form the input for an ensemble of logistic regression classifiers, achieving SOTA in the AA task. It is important to note that in (Tyo et al., 2022), the authors report macro-averaged accuracy, while in (Kestemont et al., 2018), the evaluation metric is macro-averaged F1-score. Al-

#Auth	Dataset	MFLM	Stylo	SBERT	Word	Char	SBERT+	Word+	Char+	SOTA
62	IMDB62	0.91	0.02	0.87	0.82	**0.92**	0.94	0.96	0.96	0.99*
50	PAN06	0.58	0.00	0.62	**0.65**	0.64	0.64	0.68	0.66	0.77*
20	PAN18 P1	**0.57**	0.01	0.30	0.40	0.52	0.55	0.59	0.60	0.65**
15	PAN18 P2	**0.63**	0.00	0.40	0.50	0.62	0.64	0.66	0.66	0.68**
10	PAN18 P3	0.67	0.20	0.50	0.63	**0.82**	0.67	0.71	0.74	0.74**
5	PAN18 P4	0.57	0.20	0.49	**0.69**	0.57	0.64	0.69	0.66	0.68**

Table 5: Authorship Attribution evaluation results, reporting the weighted F1-score. With bold font, we draw attention to the best performing model. Word and Char refer to the word n-grams and character n-grams respectively. Columns where the title line contains a '+' character correspond to experiments where we concatenated MFLM embeddings to the baseline document vector. *Macro-averaged accuracy as reported in (Tyo et al., 2022). **Macro-averaged F1-score as reported in (Kestemont et al., 2018).

though a direct comparison may not be feasible due to these differing metrics, these results offer valuable insight into the task's complexity.

5 Conclusion

This study investigated two research questions regarding the detection and application of Figurative Language (FL) features in machine learning.

Firstly, we explored whether a multi-task model trained to simultaneously detect multiple FL features (Metaphor, Simile, Idiom, Sarcasm, Hyperbole, and Irony) could outperform individual models specialized for each feature. By leveraging RoBERTa-Large and a multi-label training dataset derived from binary classifiers, our Multi-task Figurative Language Model (MFLM) achieved superior performance on 8 out of 13 test sets, particularly excelling in detecting Simile, Idiom, Irony, and Hyperbole. This finding highlights the increased effectiveness of a unified approach for comprehensive FL detection.

Secondly, we examined the potential of incorporating FL features to enhance performance in Authorship Attribution (AA) tasks. Utilizing three diverse AA datasets and Multi-Layer Perceptron (MLP) classifiers, we evaluated the contribution of MFLM sentence embeddings alongside various baseline features like Stylometric features, SBERT Embdeddings and word and character n-gram vectors. The results showed the competitive performance achieved by MFLM embeddings alone, while their combination with other features yielded consistent performance improvements across nearly all cases. This strongly supports the second research question, indicating the positive impact of integrating FL features in AA tasks.

Our study offers valuable insights into the effectiveness of multi-task learning for comprehensive FL detection and the potential of FL features to improve AA tasks. Further research could explore the applicability of MFLM to additional NLP tasks, such as sentiment analysis and information retrieval. Moreover, future studies could investigate the impact of incorporating additional FL features into a single classification model, such as personification, metonymy, onnomatopoeia, etc., as well as domain-specific knowledge for even more refined FL detection and application.

Acknowledgements

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200002 and the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001121C0186. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Ethical Considerations and Limitations

In this paper, we investigate the efficacy of training a multi-task classification model to detect Figurative Language (FL) features compared to specialized binary models. In this work, we also explore leveraging the multi-task model embeddings for the Authorship Attribution (AA) task.

One potential limitation of our study arises from the combination of different datasets for the various Figurative Language (FL) features under consideration. The quality of annotations across these datasets is not uniform, with some lacking annota-

tion manuals or relying on automatic and crowd-sourced approaches for dataset creation. This inconsistency can introduce errors into our model. Furthermore, the datasets, while publicly available, may contain inherent biases due to the lack of clear instructions for annotating literal sentences and potential variability in human annotator judgments.

In the process of constructing annotated corpora for training machine learning algorithms for automatic figurative language detection, it's crucial to consider the interpretive discrepancies between experts and non-experts. The annotations found in the collection of datasets used in this study are all taken as ground truth of equal importance, potentially leading towards a biased FL detection model. An expert, with their nuanced understanding, can identify subtle metaphors and idioms that may elude an ordinary reader. However, non-experts, influenced by their unique cultural backgrounds and personal experiences, may interpret figurative language differently (Carrol and Littlemore, 2020; Robo, 2020). For instance, certain phrases may have specific connotations in one culture but be meaningless in another. Similarly, a person's familiarity with a subject matter can greatly influence their understanding of related figurative language. Therefore, to ensure a more accurate and comprehensive analysis of figurative language, these factors must be taken into account. Future work will address this issue by conducting qualitative and quantitative analyses on the annotated datasets.

In our methodology, we employ specialized binary models for each feature, trained on our combined datasets, to predict figurative language labels for the training examples used in our multi-task model. This approach, while effective, can lead to error propagation, resulting in incorrect predictions from our model. However, our evaluation and manual error analysis indicate that our multi-task model's predictions are often reasonable, with errors frequently attributable to incomplete human annotations.

The second part of our study applies the embeddings from our multi-task FL model to the AA task. We train MLP classifiers using document vectors as features on three publicly available datasets, each focusing on a different topic: movie reviews, corporate/industrial topics, and fan fiction. However, these topics are not very diverse, which could introduce bias into the datasets with respect to authorship. For instance, in the fan fiction dataset, some authors may exclusively write "Harry Potter" fan fiction, which could skew the evaluation of different features.

Lastly, it is important to note that the predictions of deep neural language models, such as the ones used in our study, are often difficult to interpret and explain. This lack of interpretability is a common challenge in the field and is another limitation to consider in our work.

References

Tosin P Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaidou, Foteini Liwicki, and Marcus Liwicki. 2021. Potential idiomatic expression (pie)-english: Corpus for classes of idioms. *arXiv preprint arXiv:2105.03280*.

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Jonathan Anderson. 1981. Analysing the readability of english and non-english texts in the classroom with lix. *Non-journal publication.*

Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 388–401, Toronto, Canada. Association for Computational Linguistics.

Leonard Bauersfeld, Angel Romero, Manasi Muglikar, and Davide Scaramuzza. 2023. Cracking double-blind review: Authorship attribution with deep learning. *PLOS ONE*, 18(6):e0287611.

Rhys Biddle, Maciek Rybinski, Qian Li, Cecile Paris, and Guandong Xu. 2021. Harnessing privileged information for hyperbole detection. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 58–67, Online. Australasian Language Technology Association.

J Briskilal and C.N. Subalalitha. 2022. An ensemble model for classifying idioms and literal texts using bert and roberta. *Inf. Process. Manage.*, 59(1).

Étienne Brunet et al. 1978. *Le vocabulaire de Jean Giraudoux structure et évolution*. Slatkine.

Gareth Carrol and Jeannette Littlemore. 2020. Resolving figurative expressions during reading: The role of familiarity, transparency, and context. *Discourse Processes*, 57(7):609–626.

John B Carroll. 1964. Language and thought. *Reading Improvement*, 2(1):80.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science : Interpreting figurative language in narratives.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. Flute: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159.

J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

A Crawford. 1985. Fórmula y gráfico para determinar la comprensibilidad de textos del nivel primario en castellano. *Lectura Y Vida*, 4:18–24.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Daniel Dugast. 1979. *Vocabulaire et stylistique*, volume 8. Slatkine.

Daniel Dugast. 1980. La statistique lexicale. *(No Title)*.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137.

Ibrahim Abu Farha, Silviu Oprea, Steve Wilson, and Walid Magdy. 2022. Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in english and arabic. In *The 16th International Workshop on Semantic Evaluation 2022*, pages 802–814. Association for Computational Linguistics.

José Fernández Huerta. 1959. Medidas sencillas de lecturabilidad. *Consigna*, 214:29–32.

José Ángel González, Lluís-F Hurtado, and Ferran Pla. 2020. Transformer based contextualization of pre-trained word embeddings for irony detection in twitter. *Information Processing & Management*, 57(4):102262.

P. Guiraud. 1954. *Les caractères statistiques du vocabulaire: essai de méthodologie*. Presses universitaires de France.

R. Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

Gustav Herdan. 1955. A new derivation and interpretation of yule's 'characteristic'k. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 6:332–339.

Simona Herdan and Yael Sharvit. 2006. Definite and nondefinite superlatives and npi licensing. *Syntax*, 9(1):1–31.

Antony Honoré et al. 1979. Some simple measures of richness of vocabulary. *Association for literary and linguistic computing bulletin*, 7(2):172–177.

John Houvardas and Efstathios Stamatatos. 2006. N-gram feature selection for authorship identification. In *International conference on artificial intelligence: Methodology, systems, and applications*, pages 77–86. Springer.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018. Overview of the author identification task at pan-2018: cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al.*, pages 1–25.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and P.G. Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195.

G. Lakoff and M. Johnson. 2008. *Metaphors We Live By*. University of Chicago Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Longxuan Ma, Weinan Zhang, Shuhan Zhou, Churui Sun, Changxin Ke, and Ting Liu. 2023. I run as fast as a rabbit, can you? a multilingual simile dialogue dataset. *arXiv preprint arXiv:2306.05672*.

Heinz-Dieter Mass. 1972. Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 2(8):73.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Rachel McAlpine. 2006. From plain english to global english.

Sven Meyer zu Eissen, Benno Stein, and Marion Kulig. 2007. Plagiarism detection without reference collections. In *Advances in Data Analysis: Proceedings of the 30 th Annual Conference of the Gesellschaft für Klassifikation eV, Freie Universität Berlin, March 8–10, 2006*, pages 359–366. Springer.

R Michéa. 1969. Répétition et variété dans l'emploi des mots. *Bulletin de la Société de Linguistique de Paris*, 64(1):1–24.

R Michéa. 1971. De la relation entre le nombre des mots d'une fréquence déterminée et celui des mots différents employés dans le texte. *Cahiers de lexicologie*, 8(1):65–78.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the lcc metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2008–2018.

John O'hayre. 1966. *Gobbledygook has gotta go*. US Department of the Interior, Bureau of Land Management.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2017. Creating and characterizing a diverse corpus of sarcasm in dialogue. *arXiv preprint arXiv:1709.05404*.

Francisco Szigriszt Pazos. 1993. *Predictive Systems for Legibility of the Written Message: Perspicuity Formula*. Universidad Complutense de Madrid, Servicio de Reprografía.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Lorena Robo. 2020. Discrepancies of figurative language use reflected through cross-linguistic and intercultural differences in english and albanian language. *European Journal of Language and Literature*, 6(1):1–14.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 87–94. Springer.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Technical report, DTIC document.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Herbert S Sichel. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547.

Edward H Simpson. 1949. Measurement of diversity. *nature*, 163(4148):688–688.

George Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

Paul Thibodeau, James L McClelland, and Lera Boroditsky. 2009. When a bad metaphor may not be a victimless crime: The role of metaphor in social policy. In *Proceedings of the 31st annual conference of the cognitive science society*, volume 29, pages 809–14. Citeseer.

Yufei Tian, Nanyun Peng, et al. 2021. Hypogen: Hyperbole generation with commonsense and counterfactual knowledge. *arXiv preprint arXiv:2109.05097*.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Juhan Tuldava. 1977. Quantitative relations between size of text and size of vocabulary. *Journal of Linguistic Calculus*, pages 28–35.

Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton. 2022. On the state of the art in authorship attribution and authorship verification.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Byron C Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516.

Mengfei Yuan, Zhou Mengyuan, Lianxin Jiang, Yang Mo, and Xiaofeng Shi. 2022. stce at SemEval-2022 task 6: Sarcasm detection in English tweets. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 820–826, Seattle, United States. Association for Computational Linguistics.

C Udny Yule. 2014. *The statistical study of literary vocabulary*. Cambridge University Press.

Yunxiang Zhang and Xiaojun Wan. 2021. Mover: Mask, over-generate and rank for hyperbole generation. *arXiv preprint arXiv:2109.07726*.

A Appendix

A.1 Appendix: Figurative Language Datasets

Here we present additional details regarding our 13 figurative language datasets. In Tables 6 and 7 we show the number of examples per class label for the train/test sets of all datasets. The datasets that had predefined train/test splits are: FLUE, iSarcasm, and Irony SemEval 2018. For the remaining datasets, we reserve a 10% stratified sample for testing. In the following paragraphs we will be discussing some interesting datasets.

The only corpus in our collection that is truly multi-labeled is the iSarcasm dataset. The curators of iSarcasm created the collection by recruiting Twitter users and asking them to specify one sarcastic and three non-sarcastic tweets from their posted messages. Then, they asked the participants to provide a literal rephrase for every sarcastic message that conveys the same meaning. Furthermore, for every sarcastic message, the authors perform a second annotation stage where they further label these messages with irony, overstatement (hyperbole), understatement, satire, and rhetorical questions. In our work, we assume that the rephrases provided by the original participants are indeed literal sentences, however, we do not make the same assumption for the non-sarcastic messages that were also provided. In addition, since in our research we focus on six figurative language types, we ignore labels that are outside of this set. In such cases, we retain the sentence with only the sarcasm / not_sarcasm, irony / not_irony or hyperbole / not_hyperbole labels.

The Sarcasm Corpus is a multi-class dataset centered around the binary classification task of sarcastic sentences. However, an extension of the dataset (separate file) contains sarcastic and non-sarcastic sentences that all have hyperbole. Since this was an addition to the main corpus, we cannot assume that the remaining files are completely devoid of hyperbole. Therefore, the hyperbolic sentences also have sarcasm / not_sarcasm labels, but not the other way around.

The FLUTE dataset is also multi-class dataset, which means that each example is either metaphor, simile, sarcasm, or idiom. Each figurative sentence is paired with the two literal paraphrases, one aligning with the actual meaning of the figurative sentence, and the other communicating the opposite meaning. For instance, the figurative sentence *"After a glass of wine, he loosened up a bit"* will have a literal counterpart *"After a glass of wine, he relaxed up a bit"* and the opposite paraphrase would be *"After a glass of wine, he stressed up a bit"*.

PIE-English is another interesting dataset. Here, the authors have automatically created a collection of sentences that contain possible idiomatic expressions. With further manual annotation efforts, they annotated each sentence whether its literal, therefore not idiomatic, or the idiom is constructed using euphemism, metaphor, personification, simile, parallelism, paradox, hyperbole oxymoron, or irony. Thus, every figurative sentence is an idiom plus an other figurative language class. In this work we focus on six figurative language types, so we ignore labels that are outside of this set. In such cases, we retain the sentence with only the idiom label.

A.2 Appendix: Figurative Language Classification Binary Training Sets

To train specilized binary models to detect FL features, we merge datasets annotated with examples relevant to each specific feature. For instance, to train a classifier for metaphors, we aggregate data from PIE-English, FLUTE, LCC, and MOH datasets. Similarly, for simile classification, we gather data from PIE-English, FLUTE, MSD23, and Figurative Comparisons datasets. Table 8 shows the number of positive, negative and literal examples used to train each binary classifier.

Datasets	Meta.	Sarc.	Hyp.	Irony	Not Meta.	Not Sarc.	Not Hyp.	Not Irony
Reddit Irony Corpus	-	-	-	483/54	-	-	-	1271/141
Irony SemEval18	-	-	-	1901/311	-	-	-	1916/473
iSarcasm	-	713/133	40/6	155/19	-	3622/1059	4295/1186	4180/1173
Sarcasm Corpus	-	4223/470	1047/117	-	-	4224/469	-	-
LCC	2732/304	-	-	-	3972/441	-	-	-

Table 6: (Appendix) Class distribution between train/test sets for each dataset. These datasets do not have 'literal' annotations.

Datasets	Metaphor	Simile	Sarcasm	Hyperbole	Idiom	Irony	Literal
MOVER	-	-	-	906 / 101	-	-	1997 / 222
HypoGen	-	-	-	1688 / 188	-	-	2585 / 287
EPIE	-	-	-	-	2485 / 276	-	337 / 38
PIE-English	11330 / 1260	965 / 107	-	41 / 5	12363 / 1375	27 / 3	966 / 107
FLUTE	625 / 124	625 / 125	2216 / 461	-	884 / 125	-	6368 / 1326
MSD23	-	3218 / 358	-	-	-	-	4113 / 457
Figurative Comparisons	-	404 / 45	-	-	-	-	856 / 95
MOH	369 / 41	-	-	-	-	-	1106 / 123

Table 7: (Appendix) Class distribution between train/test sets for each dataset. These datasets have 'literal' annotations.

A.3 Appendix: Figurative Language Classification Multi-label Training Set

We use the fine-tuned specialized binary classification models to automatically tag our training corpora in a multi-label format. Table 9 shows the number of examples per figurative language class, as predicted by the binary classifiers. This dataset forms the basis of training our multi-task model. At a later step, this dataset gets split in train/dev set, where a 10% stratified sample is reserved for development.

A.4 Appendix: Figurative Language Classification Error Analysis

In this subsection of the appendix, we present additional randomly selected examples where the model MFLM and binary model predictions do not align with human annotations. These additional examples are presented in Table 10.

A.5 Appendix: Authorship Attribution Baselines

In this section of the appendix we provide further details regarding the Stylometric features of our Authorship Attribution (AA) baseline approach. We implement 52 text metrics using the cophi[6] and textstat[7] Python packages. These metrics are used to form a document vector with 52 stylometric features. In the Table 11 we list the feature names along with implementation notes.

[6]cophi: https://github.com/cophi-wue/cophi-toolbox

[7]textstat: https://github.com/textstat/textstat

Classifier	Positive	Negative	Literal
Metaphor	15056	3972	11084
Simile	5212	0	5212
Sarcasm	7152	3576	3576
Hyperbole	3576	1861	1861
Idiom	15732	0	15732
Irony	2566	1283	1283

Table 8: (Appendix) The number of positive, negative and literal examples used to train each binary classifier.

Metaphor	Simile	Sarcasm	Hyperbole	Idiom	Irony
18981	6618	9906	13699	18604	10176

Table 9: (Appendix) Class distribution for the combined multi-labeled training dataset.

MFLM & Binary models disagree with GT

GT	Literal	*Stupidity was **as important as** intelligence, and as difficult to attain.*
MFLM	Simile	
Bin	Simile	
GT	Literal	*This office is **as lively as** a bustling beehive.*
MFLM	Simile, Hyperbole	
Bin	Simile, Hyperbole	
GT	Literal	*They decided to continue, but within five minutes Sustad broke an ice hammer, **forcing them to retreat in mockingly perfect weather**.*
MFLM	Sarcasm, Hyperbole	
Bin	Sarcasm, Hyperbole	
GT	Not Irony	*The letter and article seem to speak more of John Boehner wanting to **fire a gut** for criticizing the Pope. Misleading title.*
MFLM	Idiom, Sarcasm, Irony	
Bin	Idiom, Irony	

MFLM disagrees with GT, Binary model agrees with GT

GT	Literal	***I ace through** the work.*
MFLM	Metaphor	
GT	Not Sarcasm, Not Irony	***Full throttle**? 11 players changed and playing the philosophy that the manager wants isn't grounds for slagging! Especially when we win! **Clutching at straws** here!*
MFLM	Metaphor, Idiom, Irony	
GT	Literal	*This **dirty money** we're using to finance the campaign is a risk!*
MFLM	Metaphor	
GT	Literal	*The leaves clog our drains in the Fall*
MFLM	Metaphor	
GT	Not Metaphor	*If you are trying to claim gun control is not incremental I am first going to **laugh my head off** at such an obviously stupid statement.*
MFLM	Metaphor, Irony, Hyperbole	

Table 10: (Appendix) Samples where model predictions do not align with human annotations.

Feature Name	Notes
Average Word Length Chars	Average number of characters per word.
Average Syllables Per Word	Average number of syllables per word.
Average Sentence Length	Average number of words per sentence.
Average Sentence Length Chars	Average number of characters per sentence.
Average Word Frequency Class	(Meyer zu Eissen et al., 2007)
Type Token Ratio	Number of unique words (types) over the total number of words (tokens).
Digit Ratio	Number of numerical characters over total number of characters.
Puncuations Ratio	Number of punctuation characters over total number of characters.
Uppercase Ratio	Number of uppercase letter characters over total number of characters.
Special Characters Ratio	Number of special characters over total number of characters.
Stopword Ratio	Number of stopwords over total number of words.
Functional Words Ratio	Number of functional words over total number of words.
Hapax Legomena Ratio	Number of words that appear once over total number of words.
Hapax Dislegomena Ratio	Number of words that appear twice over total number of words.
Automated Readability Metric	(Senter and Smith, 1967)
Flesch Reading Ease Metric	(Kincaid et al., 1975)
Flesch Kincaid Grade Metric	(Kincaid et al., 1975)
Dale Chall Readability Metric	(Dale and Chall, 1948)
New Dale Chall Readability Metric	(Chall and Dale, 1995)
Spache Readability Metric	(Spache, 1953)
Gunning Fog Metric	(Gunning, 1952)
Lix Index	Average sentence length plus the percentage of words of more than six letters.
Rix Index	(Anderson, 1981)
Fernandez Huerta Index	(Fernández Huerta, 1959)
Szigriszt Pazos Index	(Pazos, 1993)
Crawford Index	(Crawford, 1985)
Mcalpine Eflaw Metric	(McAlpine, 2006)
Guiraud R Metric	(Guiraud, 1954)
Herdan C Metric	(Herdan and Sharvit, 2006)
Dugast K Metric	(Dugast, 1979)
Maas A2 Metric	(Mass, 1972)
Dugast U Metric	(Dugast, 1980)
Tuldava LN Metric	(Tuldava, 1977)
Brunet W Metric	(Brunet et al., 1978)
Corrected Token Type Ratio	(Carroll, 1964)
Summer S Index	Similar to TTR, $S = log(log(\text{types}))/log(log(\text{tokens}))$.
Sichel S Metric	(Sichel, 1975)
Michea M Metric	(Michéa, 1969, 1971)
Honore H Metric	(Honoré et al., 1979)
Shannon Entropy	(Shannon, 1948)
Yule K Metric	(Yule, 2014)
Simpson D Metric	(Simpson, 1949)
Herdan VM Metric	(Herdan, 1955)
Coleman Liau Metric	(Coleman and Liau, 1975)
Linsear Write Metric	(O'hayre, 1966)
Smog Metric	(Mc Laughlin, 1969)
Threshold Word Length H Ratio	Number of words with more than 5 characters over total number of words.
Threshold Word Length L Ratio	Number of words with less than 5 characters over total number of words.
Threshold Syllables Per Word H Ratio	Number of words with more than 2 syllables over total number of words.
Threshold Syllables Per Word L Ratio	Number of words with less than 2 syllables over total number of words.
Threshold Sentence Length H Ratio	Number of sentences with more than 17 words over the total number or sentences.
Threshold Sentence Length L Ratio	Number of sentences with less than 17 words over the total number or sentences.

Table 11: (Appendix) List of Stylometric feature names and implementation notes.