Hierarchical Embedding Topic Modeling for Stance Detection on Unknown Targets

António Câmara*

10 May 2022

Abstract

Stance detection is an emerging problem in natural language processing with broad application to the social sciences that seeks to understand how authors express attitudes. However, existing models and datasets are only developed for settings where the stance object, or topic of debate, is known. Moreover, existing settings for this problem are high-resource and do not consider the relationship between topics. In this paper, we introduce a novel task, stance detection on unknown targets, that seeks to measure a model's ability to detect stance on topics discovered using only the text itself. To that end, we introduce a model that first discovers a hierarchical set of topics for stance detection using a semantic embedding space and then uses large-scale transformer-based language models for stance detection on these discovered topics. In comparison to popular models, we find that our model performs well on topic modeling, stance detection, and our novel task, especially in low-resource and hierarchical settings. We also discuss the application of our work in low-resource settings and begin collecting datasets that study African American English and Black communities online.

*B.S. Candidate in Computer Science, Columbia University. ac4443@columbia.edu

Contents

List of Tables

List of Figures

Acknowledgements

I am grateful for the support of Professor Kathleen McKeown and Professor Desmond Patton on this research project. I am especially grateful for the amazing mentorship and guidance of Emily Allaway and Elsbeth Turcan. This thesis would not have been possible without the love and support of my family and friends. In particular, thank you to Ana Câmara, Isabel Câmara, Olivia Hussey, and Max Helman.

1 Introduction

Stance detection is an emerging problem in natural language processing that seeks to classify the *stance*, or the linguistic representation of attitude, of the author of a document on a particular *target*, or stance object. Stance detection is an important task for both computer scientists and social scientists as it is critical in understanding how actors express attitudes on social and political issues and in learning how implicit knowledge is expressed in text. This is because an author may express stance across a variety of targets that may or may not be mentioned explicitly in a document. At the same time, such documents help to inform our understanding of an actor's beliefs, values, and understanding of the world. For example, consider the following document from Allaway and McKeown (2020):

> "every day police officers take advantage of their authority and act in ways which they would not act if they were on camera."

In this document, we consider two targets: the mandatory use of body cameras by police officers (`body cameras`) and the institution of policing (`police`). It is clear the author is against, or is `CON` on, `police` due to their belief that police officers "take advantage of their authority" on a regular basis. However, the author also expresses support, or is `PRO` on, `body cameras` since body cameras may serve as a policy intervention that would prevent abuse by police. In this example, there is a logical relationship between an author expressing a negative opinion of policing and a positive opinion of body cameras because such positions are related. An author who is against policing because they believe that police are abusive are also likely to support interventions that are perceived to reduce abuse in policing. However, detecting such stances and relationships between stances are difficult for a computer to learn. This is because such a task requires a model to possess both implicit and domain knowledge in natural language.

In part due to its recent development and task difficulty, stance detection models and datasets are typically developed for high resource settings where targets are both known and the relationship between targets is not considered. However, performing stance detection in low resource settings is of particular interest to researchers because such settings permit for the study of communities that either speak a minority language or dialect, constitute a minority community with targets of interest that are unknown to researchers, or are otherwise underrepresented in political, social, academic, or technical spaces in ways that preclude their study. Alongside cases where the salient targets of an understudied community are not known to researchers, developing stance detection methods for data that does not include explicit targets is essential for the advancement and expansion of stance detection into related disciplines, such as the social sciences. Finally, harnessing information about the relationships between targets is critical, especially in low resource settings, for gaining a more complete understanding of how the attitudes of authors are related both linguistically and ideologically.

In this paper, we study these limitations of stance detection and harnessing topic modeling to understand how people express stance in unknown, complex, and low-resource settings. Our contributions are summarized as:

- We introduce a novel task, stance detection on unknown targets, for settings where stance targets are unknown or unavailable.
- We introduce two models: the Hierarchical Embedding Topic Model (`HETM`) and the BERT-powered Stance Detector (`BERT-SD`) for topic modeling and stance detection respectively.
- We perform experiments on these models, alongside baselines, in three tasks: topic modeling, stance detection on known targets, and stance detection on unknown targets.

- We collect sample datasets for the study of Black communities and African American English in natural language processing.

The rest of the paper is organized as follows. First, we provide an extensive literature review of the stance detection and topic modeling literature. Second, we discuss our efforts in creating stance detection datasets in African American English and for the study of Black communities as well as discuss the datasets we use for our experiments in topic modeling and stance detection. Third, we introduce the Hierarchical Embedding Topic Model (`HETM`), the BERT-powered Stance Detector (`BERT-SD`), and discuss their application in performing topic modeling and stance detection in unknown, complex, and low-resource settings. Fourth, we evaluate and discuss the performance of our models and popular baselines in topic modeling and stance detection tasks. Fifth, we consider several future directions as well as the ethical considerations of our work.

2 Related Works

2.1 Stance Detection

In computer science, stance detection is a sentiment analysis task that seeks to classify an author's stance on a stance object, or target. Such a target may be a concept or topic, represented by a short noun phrase (e.g. `states' rights`) or an idea or position, represented by a statement (e.g. `abortion should be federally legalized`). A target may be explicitly or implicitly referenced in the document; however, existing datasets always include targets alongside documents such that classification is performed on the pair itself (Allaway and McKeown, 2020). Formally, given a document d and a target t, we classify the stance of d on t either in the affirmative, `PRO` for phrases and `FAVOR` for statements, or in the negative, `CON` for phrases and `AGAINST` for statements. Additional classes are also present in stance datasets,

such as NEUTRAL, where no stance is expressed, and UNRELATED, where stance is not expressed on the corresponding object. The exact class names differ across datasets (Küçük and Can, 2020). However, since the target may go left unsaid or since the detection of stance itself may rely on real-world domain knowledge, harnessing implicit knowledge is key in building successful systems. Moreover, a document may express stance at multiple, potentially related targets. In multi-target stance detection, a single document has multiple targets such that each unique document-target pair is classified. Earlier predictions for a given document may also inform future predictions for that same document (Sobhani et al., 2017; Allaway and McKeown, 2020). This is because targets may be either logically related or interconnected through broader frameworks such as ideology. In cross-target stance detection, models learn to classify stance for targets without training data using transfer learning for different albeit related targets (Augenstein et al., 2016; Xu et al., 2018). In practice, there may be limited training data for each target. Two related tasks are zero-shot and few-shot stance detection where classification takes place for targets with no and few training examples respectively (Allaway and McKeown, 2020). In this paper, we are interested in expanding low-resource stance detection to settings beyond limited or no training data to settings where targets themselves are not known a priori and must be learned alongside the stance detection task itself.

In linguistics, stance is best formalized by Du Bois (2007) as a public communication that simultaneously evaluates an object, declares a position in conversation with one's sociocultural context, and aligns oneself with likeminded actors. The concept of stance is also closely related with other concepts in the social sciences. For example, stance is similar to concepts of attitude and opinion in psychology and political science. While definitions vary, an attitude is best understood as an affect towards or evaluation of an attitude object that is both informed by one's beliefs on or about the object (Fishbein and Ajzen, 1972). Opinions are expressions of attitude in natural

language. In the social sciences, opinions are used to study attitudes since attitudes are latent and thus immeasurable (Erikson and Tedin, 2015). The evaluation of an object, or target, is a critical component of detecting stance, describing opinion, and understanding attitude. Moreover, the social contextualization and alignment that helps constitute stance is closely related both to group identification and participation in public life. As such, stance detection is a powerful tool for the study of public opinion. Public opinion is an important area of research in political science with applications in interpreting the policy demands of the public and forecasting the outcomes of elections and referenda (Erikson and Tedin, 2015). While the study of public opinion is central to computational social science, an emerging field that seeks to understand and explain the social world utilizing computational methods and large-scale datasets, limitations of existing stance detection work preclude further study. For example, existing stance detection datasets and models include and require the stance object, or target, during training and thus are limited to settings where such targets are known and available.

Stance detection is broad in its genre, methodology, and further application. For example, existing work in stance detection has ranged from highly standardized settings, such as the U.S. Congressional Record (Thomas et al., 2006), to more informal settings, such as social media services (Mohammad et al., 2016) and online debate forums (Gottopati et al., 2013). The methodological approaches to stance detection are also varied. Traditionally, a wide suite of supervised machine learning algorithms, including support vector machines and logistic regression, are utilized and compared in studies, sometimes employing ensembles for ultimate predictions (Küçük and Can, 2020). However, deep learning methods such as recurrent neural networks (RNNs), convolutional neural networks (CNNs) (Augenstein et al., 2016), and methods that harness pretraining on large corpora to improve natural language understanding such as `fastText` (Mikolov et al., 2018) and `BERT` (Devlin et al., 2018) are increasingly

common. Stance detection has been applied to a broad range of related tasks. For example, stance detection has been employed in improving the state of the art in argument mining, sarcasm detection, and fake news detection (Küçük and Can, 2020).

While stance detection is an important question with broad application to the social sciences, existing models and datasets are relatively limited in their scope and utility; however, some work has been done to expand stance detection into low-resource and more complex settings. Many popular datasets in stance detection are limited to a small number of targets that are narrow in scope. For example, SemEval-2016 Task 6: Detecting Stance in Tweets (`SEMEVAL`) is a popular stance detection dataset introduced by Mohammad et al. (2016). However, the dataset only includes five targets with training data and one target without training data, with all targets focusing on contemporary American politics. Even datasets that include a relatively high number of targets, such as Vamvas and Sennrich (2020), are still limited in scope to targets dependent on the genre and source of the underlying data and constrained to those targets which were identified as sufficiently salient by annotators. Given the disproportionately low representation of women, people of color, and other minorities in computer science and natural language processing, limitations in stance detection requiring manual annotation of targets are likely to neglect the study of those minorities. For example, to the best of our knowledge, there exist no stance detection datasets in African American English or that focus on targets of interest to Black communities.

While no stance detection model has yet to be developed for settings where targets are not known, some work has been done to detect stance on targets with limited or no training data. For example, Allaway and McKeown (2020) introduces the Varied Stance Topic (`VAST`) dataset, which includes diverse targets ranging from American politics to international affairs to debates in religion and public health. Moreover, multiple targets express the same concepts using different frames and expressions to

better reflect typical parlance. Likewise, some work has been done on multilingual settings. Vamvas and Sennrich (2020) introduces a cross-lingual stance detection dataset in German, French, and Italian across 150 targets regarding Swiss politics while Taulé et al. (2017) introduces a multilingual dataset in Spanish and Catalan on targets concerning Catalan independence. However, these datasets nevertheless focus on relatively high-resource European languages and issues salient to European communities. Moreover, no dataset to our knowledge describes relationships between targets. Although the Debatepedia (`DEBATEPEDIA`) dataset introduced by Gottopati et al. (2013) does label documents with both a supertarget and a subtarget, the supertargets are not appropriate for stance detection because they are too broad (e.g. `war`, `politics`).

In settings where targets have limited or no training data, understanding the relationships between targets is critical in detecting stance. This is because knowledge learned by a model for high-resource targets may be transferred in service of understanding low-resource targets. For multi-target stance detection, Hasan and Ng (2013) trains an independent and unique classifier for each target while Sobhani et al. (2017) introduces a model that can simultaneously classify the stance of a document on two targets as well as a framework for simultaneous classification on any number of targets. Augenstein et al. (2016) uses the `SEMEVAL` dataset and bidirectional LSTM encodings of documents conditional on targets to both detect stance in settings where the target is implicit and in settings where the there is no training data for a target. Xu et al. (2018) also employs the `SEMEVAL` dataset and bidirectional LSTMs but introduces a self-attention mechanism to perform cross-target stance detection on domain-related targets (e.g. targets in the domain of American politics include `Hillary Clinton` and `Donald Trump`). On the other hand, Vamvas and Sennrich (2020) uses a large-scale pretrained language model, multilingual `BERT`, to perform zero-shot, cross-lingual, and cross-target stance detection while Allaway and McKe-

own (2020) uses both BERT and clustering for generalized target representations to perform zero-shot stance detection.

2.2 Topic Modeling

We turn to the literature on topic modeling, which we utilize in service of automatic target discovery for stance detection. Topic modeling is a task in machine learning that seeks to discover a set of latent variables that best characterize a collection of data points. In natural language processing, this implies discovering a set of topics that best describes a corpus (Vayansky and Kumar, 2020).

Latent Dirichlet allocation (LDA), introduced by Blei et al. (2003), is a canonical topic model which serves as the foundation for the models we explore further. LDA formally defines a topic as a probability distribution over a vocabulary; however, since documents are composed of words, LDA also defines a document as a probability distribution over topics. The motivation behind LDA is that we can define a generative story, or process, by which documents are randomly generated using the discovered topics. We first select a fixed number of k topics. In the generative story, we assume a collection of M documents each of N length. First, we sample k probability distributions β_j over the vocabulary, one for each topic $j \in \{1, ..., k\}$, from a Dirichlet distribution parametrized by η. Then, for each document in $i \in M$, we sample a probability distribution θ_i over the k topics from a Dirichlet distribution parametrized by α. Finally, for each word $w \in N$ in document i, we first sample a topic assignment conditional on the document's distribution over topics using a categorical distribution over the document's topic distribution $z_{wi} \mid \theta_i \sim \text{Cat}(\theta_i)$ and next we sample a word conditional on both the topic assignment for that word and the topic's probability distribution over the vocabulary $x_{wi} \mid \beta, z_{wi} \sim \text{Cat}(\beta_{z_{wi}})$. We can then reverse the generative story to infer the latent topics given our corpus. We do so by finding the posterior probability of the topic distributions over the vocabulary,

document distributions over the topics, and word assignment distributions over the document distributions, all given the observed data — the words in the documents $p(\beta, \theta, z \mid x)$. This quantity in turn depends on the hyperparameters of the Dirichlet distributions of the topics over the vocabulary and documents over the topics, α, η. In practice, we perform variational inference to maximize this posterior probability by minimizing KL divergence on the variational family and the posterior itself. In other words, we attempt to maximize the probability of generating the actual corpus from scratch. As such, `LDA` has diverging goals: to create concise documents by assigning words to as few topics as possible and to create concise topics by placing weight on as few terms as possible in each topic.

Since we are interested in modeling topics in both complex and low-resource settings, we consider three topic models that build on `LDA`. For modeling topics in complex settings, we utilize Hierarchical latent Dirichlet allocation (`hLDA`) while for modeling topics in low-resource settings, we utilize the Embedding Topic Model (`ETM`). We then consider the Tree-Structured Neural Topic Model (`TSNTM`) which combines concepts from `hLDA` and the `ETM` to model topics in complex and low-resource settings.

Hierarchical latent Dirichlet allocation (`hLDA`), introduced by Griffiths et al. (2003), is an extension to `LDA` that explicitly models the relationship between topics by modeling such topics as nodes in a tree. However, such nodes are still probability distributions over a vocabulary. Moreover, unlike `LDA` which requires a fixed number of topics k to be determined before training, `hLDA` allows for the number of topics to change as new data are introduced using a nonparametric prior — the nested Chinese restaurant process (nCRP). The original Chinese restaurant process is a mechanism to partition integers into a variable number of collections; the analogy being that a steady stream of customers enter a restaurant with infinite tables of infinite size, choosing whether to sit at one of the occupied tables or at an new unoccupied table. This allows us to model uncertainty over the number of topics in a topic model. The

nCRP therefore allows us to model uncertainty over the number of L-level trees in a topic model where each tree consists of topic-representing nodes. The restaurant analogy states that, for L nights, a customer visits a restaurant where on each table lies a menu for the restaurant the customer is to visit the next night. The use of the nCRP allows us to model topics hierarchically, since topics are represented as nodes in a tree, and nonparametrically, since the number of topics is uncertain, dependent on a hyperparameter γ and the documents considered. In this model, each additional level down the tree provides further specificity in the probability distributions over the vocabulary such that parents are not merely summaries of their children, but rather representations of the shared vocabulary of their children. One restriction of the model is that documents may only mix over topics on a single path down the tree.

The Embedding Topic Model (`ETM`), introduced by Dieng et al. (2020), differentiates itself from `LDA` by defining a topic as a point in a low-dimensional semantic space instead of as a probability distribution over a vocabulary. Formally, we define a L-dimensional word embedding matrix over a vocabulary V, $\rho \in \mathbb{R}^{L \times V}$, as well as k topics, which we embed in the same space as the word embeddings, $\alpha \in \mathbb{R}^{L \times K}$. Like `LDA`, we define a generative story which we then reverse to maximize the posterior distribution over the word and topic embeddings using variational inference. In the `ETM`'s generative story, we first draw k topics. Next, for each document d, we sample a distribution over topics θ_d from a logistic normal distribution. Finally, for each word n in the document d, we sample a topic assignment $z_{dn} \sim \text{Cat}(\theta_d)$ and a word $w_{dn} \sim \text{Softmax}(\rho^T \alpha_{z_{dn}})$. Note that the Softmax converts the topic embedding into a probability distribution over the vocabulary. Moreover, we utilize a neural network to sample distributions over topics, parametrized by ν. As such, we may either learn both topic and word embeddings during training or plug in pretrained word embeddings such as `GloVe` (Pennington et al., 2014). The `ETM` notably performs strongly in

a metric known as topic quality, the product of topic cohesion, a measure of interpretability, and topic diversity, a measure of uniqueness. While the authors claim that the ETM is especially apt for handling settings with large vocabularies and heavy tails, we believe that through harnessing pretrained word embeddings, we can learn richer topics in low-resource settings as well. This is because such embeddings inherently encode knowledge that can then be exploited during training by the embeddings to better and more easily represent topics in the text.

The Tree-Structured Neural Topic Model (TSNTM), introduced by Isonuma et al. (2020), combines concepts from both hLDA and ETM. TNSTM models topics in both a hierarchical and nonparametric setting using nodes in a dynamically-size tree. These nodes are implemented as hidden states in an RNN with two recurrences such that topics are embedded in a low-dimensional space as opposed to represented as probability distributions over vocabularies. TSNTM achieves improved performance in topic interpretability over hierarchical models that harness nCRP and improves scalability for large corpora. Since our model in part re-implements the general architecture of the TSNTM, we discuss this model at length in the model section (§4.1). We utilize these models in service of discovering latent topics that characterize our corpora. We then hope to generate targets, or stance objects, by sampling phrases or collections of words from these learned topics.

3 Data

3.1 Black Communities Datasets

We performed exploratory data collection and management in service of an NSF grant for the study of sentiment and emotion expression in Black communities. Due to resource and human limitations, we were tasked with discovering existing datasets or creating novel datasets using social media service data that were either in African

American English or otherwise characterized the Black community. We considered two approaches: Twitter and Reddit.

For our Twitter approach, we first collected large-scale existing datasets of tweets. Since Twitter only permits the publication of tweet IDs instead of the raw text and metadata, we first "hydrate", or retrieve tweet information from tweet IDs. We hydrate tweets from the following datasets: `COVID`, a continuously updated dataset consisting of multilingual tweets relating to COVID-19 introduced by Chen et al. (2020); `COVAXXY`, a continuously updated dataset consisting of English-language tweets relating to COVID-19 vaccines introduced by DeVerna et al. (2021); `AVAX`, a continuously updated dataset consisting of English-language tweets relating to vaccine hesitancy, introduced by Muric et al. (2021); `BLM`, a dataset collected from 2013-2020 consisting of multilingual tweets including the keywords `BlackLivesMatter`, `AllLivesMatter` and `BlueLivesMatter` introduced by Giorgi et al. (2020); and `ELECTION`, a dataset collected from 2019-2021 consisting multilingual tweets relating to the 2020 U.S. presidential election introduced by Chen et al. (2021). Since the datasets are extremely large, oftentimes composed of billions of tweets, we were only able to hydrate a small fraction of the tweets from any given dataset for development purposes before this direction of work was abandoned.

We next built a tool that takes a tweet with geolocation information and matches its coordinates to its 2020 U.S. Census tract, the smallest area of enumeration for which race and ethnicity information is publicly available. We then filtered our collected datasets for tweets that were authored in either plurality Black, majority Black, or 99% Black Census tracts. We perform this procedure on the following datasets: SemEval-2016 Task 6: Detecting Stance in Tweets, (`SEMEVAL`), a stance detection dataset on targets about politics, `COVAXXY`, and `AVAX`. `COVID`, `BLM`, and `ELECTION` were not sufficiently hydrated when we conducted this experiment. However, there were major flaws with this method. First, the overwhelming majority of tweets are not

geolocated, limiting the size of any filtered dataset. For example, on `SEMEVAL`, only 214 tweets for a single target, `atheism`, were geolocated to be from Black-majority Census tracts. Second, a tweet originating from a given Census tract does not imply that the author of that tweet is a resident of that tract nor that the author is racially representative of the residents of that tract. As such, it is difficult to conclude that geolocating tweets to Black-majority or Black-plurality Census tracts implies that such tweets depict Black communities or are written in African American English.

For our Reddit approach, we focus on `r/blackpeopletwitter`, or `r/bpt`[1]. `r/bpt` is a subreddit, or community, that consists of screenshots of posts authored by Black people from social media services, primarily Twitter. Users then discuss the post and its broader context in the thread, or comment section. The subreddit states in its masthead:

> Black culture has a unique way of examining the everyday and we are here to showcase that

While this setting alone would have provided ample opportunity to study Black communities online, `r/bpt` also allows users to apply for a badge displayed next to their username that identifies a user as Black. Users apply by submitting a photograph of their forearm alongside a handwritten note including their username to the moderators of the subreddit. Moreover, some threads on sensitive issues, as determined by the moderator staff, restrict participation to users who are Black or other "non-white [people of color]" according to the moderator-verification and self-identification Black badge system[2]. These threads, known as Country Club threads, alongside the Black badge system provide us with an accurate and ethical opportunities to study Black communities online.

[1] www.reddit.com/r/BlackPeopleTwitter

[2] www.reddit.com/r/BlackPeopleTwitter/comments/gumxuy/what_is_bpt_country_club_and_how_do_i_get

Category	Description
subreddit	the subreddit in which a post was posted (e.g. `r/bpt`)
post_id	a unique integer that identifies a post
author_id	a unique integer that identifies a user, the author of the post
author_username	the username of the author of the post
author_bipoc	a indicator variable active if the post author is verified to be Black
post_bipoc	a indicator variable active if the post has been designated a "country club thread"
post_title	the title of the post
post_text	the body text of the post (optional)
post_url	the link of the post
post_score	the score of the post
post_ratio	the ratio of upvotes to downvotes of a post
datetime	the utc time of the upload of the post

Table 1. Categories and descriptions for the dataset collected from the all-time highest-scoring posts from `r/bpt`

Category	Description
subreddit	the subreddit in which a post was posted (e.g. `r/bpt`)
post_id	a unique integer that identifies a post
comment_id	a unique integer that identifies a comment given a post
author_id	a unique integer that identifies a user, the author of the comment
author_username	the username of the author of the comment
author_bipoc	a indicator variable active if the comment author is verified to be Black
comment_text	the body text of the comment
comment_score	the score of the comment
comment_parent	the comment_id of the parent comment
comment_top_level	an indicator variable active if the comment is top level comment
datetime	the utc time of the upload of the post

Table 2. Categories and descriptions for the dataset collected from the comments of the all-time highest-scoring posts from `r/bpt`

We collect a small dataset for use by linguists based on the all-time highest rated posts. We also develop a scraper which may be set to run daily or collect data over a user-defined period of time. We collect data for 148 posts and 102083 comments as described in Tables 1 and 2 respectively with additional summary statistics in Table 3.

Posts focus on a wide array of issues in Black and American life such as elections, public health emergencies, and civil rights demonstrations. Moreover, posts express stance on these issues through narrative techniques such as irony and sarcasm as well as sentiment and emotion expression. We provide examples of two Country Club posts authored by Black users in Figure 1 and Figure 2. However, since the image attached to a post usually contains the lion's share of the author's message, computer

	Total	Black Author	Country Club
Posts	148	31	90
Comments	102083	6447	41204

Table 3. Summary statistics for the dataset collected from the all-time highest-scoring posts from `r/bpt`

Figure 1. Definitely masking up post-COVID

Figure 2. This is white privilege at its finest

vision techniques such as optical character recognition are necessary to meaningfully utilize this dataset.

3.2 Stance Detection Datasets

In this section, we describe the stance detection datasets we employ in evaluating our model and baselines on both topic modeling and stance detection tasks. We use three popular datasets for stance detection: `SEMEVAL`, introduced by Mohammad et al. (2016); `VAST`, introduced by Allaway and McKeown (2020); and `DEBATEPEDIA`, introduced by Gottopati et al. (2013).

First, we provide summary statistics on the datasets and their train, development, and test splits in Table 4. Note that for datasets without a three distinct splits, we create the necessary splits and ensure that any test split includes unseen targets at

each level in the hierarchy. Second, we briefly discuss each dataset individually.

SEMEVAL is one of the most commonly utilized datasets in stance detection. Introduced for the SemEval-2016 Task 6: Detecting Stance in Tweets competition, SEMEVAL consists of a relatively small number of English language tweets each annotated with a single target. The set of targets is also small, consisting of six targets broadly related to American politics and culture (Atheism, Climate Change is a Real Concern, Feminist Movement, Hillary Clinton, Legalization of Abortion, Donald Trump). The first five targets are present in the training and test set whereas the last target, Donald Trump, is an unseen target, present only in the test set. There is no formal relationship or hierarchy between targets in the dataset; however, targets are clearly related to each other. For example, Donald Trump and Hillary Clinton were both candidates in the 2016 U.S. presidential election. This dataset consists of three classes: FOR, AGAINST, and NONE. This dataset is unique for its genre: informal, non-standard English.

DEBATEPEDIA is another popular stance detection dataset. The dataset is sourced from a now defunct online debating website, Depbatepedia, where users argued about various controversial topics of interest. This dataset formally models a hierarchical relationship between targets: each document includes one supertarget and one subtarget. Supertargets are diverse in scope, ranging from American politics and International politics to religion and sports, with several subtargets belonging to each supertarget. However, supertargets are not appropriate stance targets because they are insufficiently specific. For example, it is unclear what a PRO stance on politics means whereas a PRO stance on gun control is clearly in favor of restricting access to firearms. In other words, a supertarget is a more generic category under which several subtargets fall. For example, gun control, abortion, and legalization of marijuana are all subtargets that fall under the supertarget politics. As such, we only detect stance on subtargets; however, we ensure that the

Dataset	SEMEVAL			DEBATEPEDIA			VAST		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Documents	2331	583	1956	4903	1226	1892	13072	1837	2523
Unique Docs	2331	583	1956	4903	1226	1892	1845	682	786
Supertargets	-	-	-	45	39	47	88	47	48
Subtargets	5	5	6	341	281	327	3823	316	590

Table 4. Summary statistics for the stance detection datasets

test set includes both unseen supertargets and subtargets for zero-shot evaluation. This dataset consists of four classes: `yes`, `no`, `pro`, and `con`.

`VAST` is the newest stance detection dataset we utilize. The dataset is sourced from the comment section of articles on *the New York Times*' "Room for Debate" section. We modify this dataset to formally models a hierarchical relationship between targets: we use the stance annotation from the Argument Reasoning Comprehension (ARC) Corpus (Habernal et al., 2018) as a supertarget and the stance annotations from Allaway and McKeown (2020) as a subtarget. Targets consider a broad range of issues and debates in primarily American politics and life. Supertargets include `education`, `california`, `israel`, and `politics` whereas subtargets include `the status of public teachers`, `beach access`, `palestinian occupation`, `capital gains taxes`. As is the case in `DEBATEPEDIA`, we do not detect stance on supertargets because supertargets are too broad. Note that unlike the other datasets we consider, a single document may correspond to multiple targets and multiple targets may reflect the same concept. This allows for a more realistic and robust understanding of the relationship between documents and targets since in practice, documents may take a stance on multiple, potentially related targets and individuals may conceptualize and refer to the same targets in different ways. This dataset consists of two classes: `1` and `0`.

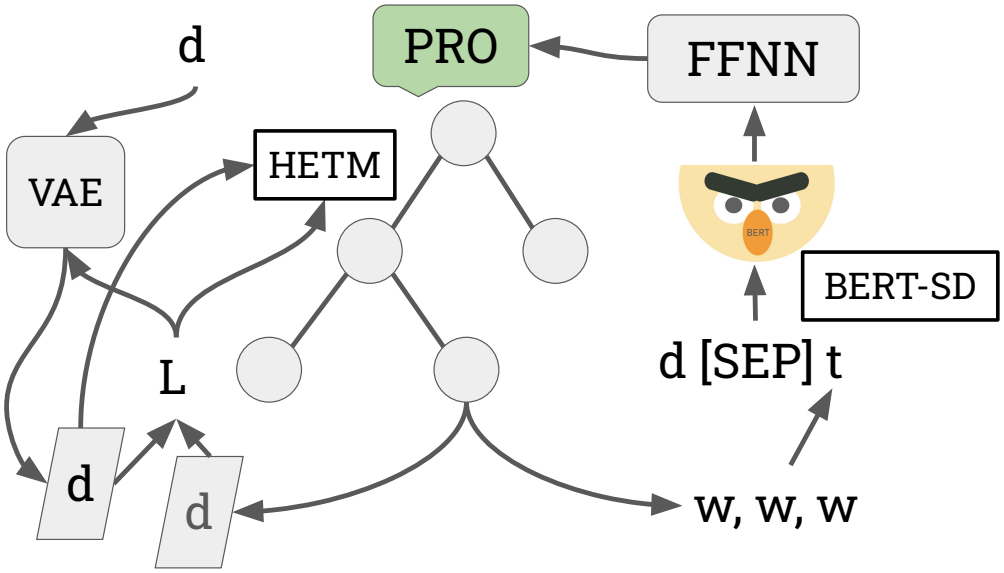

Figure 3. Architecture, `BERT` icon from Allaway and McKeown (2020)

4 Model

In this section we describe the architecture introduced in this thesis, consisting of two component models: the Hierarchical Embedding Topic Model (`HETM`) and the BERT-powered Stance Detector (`BERT-SD`).

4.1 Hierarchical Embedding Topic Model

The Hierarchical Embedding Topic Model (`HETM`) is an implementation of the `TSNTM` in PyTorch (Paszke et al., 2019) with several modifications. Recall that the `TSNTM` models topics in a hierarchical, nonparametric, and low-dimensional setting by representing topics in a growing and shrinking tree where each node is a hidden state of a doubly recurrent neural network (`DRNN`). A `DRNN` is two separate recurrences combined in a cell to generate both a hidden state and an output. One recurrence is ancestral, from a parent node to its children nodes, whereas the other is fraternal, from one node to its next sibling, if any siblings are present (Alvarez-Melis and Jaakkola, 2016). As

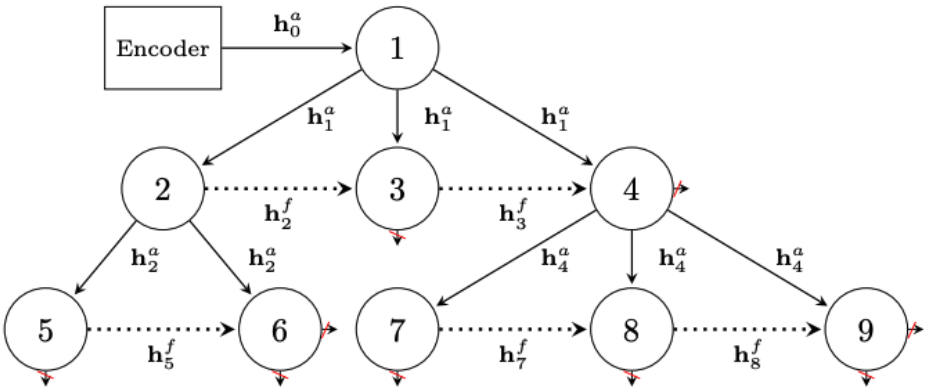

Figure 4. "Doubly" Recurrent Neural Network

such, the hidden state h_k of a topic k can be expressed by Equation 1.

$$h_k = tanh(W_p h_{par(k)} + W_s h_{k-1}) \tag{1}$$

In W_p is the ancestral recurrent weight, $h_{par(k)}$ is the hidden state of the parent of topic k, W_s is the fraternal recurrent weight, and h_{k-1} is the hidden state of the previous sibling. A diagram of a `DRNN` is provided in Figure 4

The generative process behind this model first characterizes a tree of topics. Next, for each document d, a Gaussian vector x_d is drawn. A distribution over paths down the tree π_d and a distribution over levels in the tree θ_d are sampled from neural architectures. Finally, for each word n in d, `TSNTM` samples a path $c_{d,n} \sim \text{Mult}(\pi_d)$, a level $z_{d,n} \sim \text{Mult}(\theta_d)$, and a word $w_{d,n} \sim \text{Mult}(\beta_{c_{d,n}[z_{d,n}]})$, where $\beta_{c_{d,n}[z_{d,n}]}$ is a distribution over the vocabulary assigned to the node in the $c_{d,n}$ path at the $z_{d,n}$ level. Since documents mix over all paths and levels, documents mix over all topics in the tree, unlike the `hLDA` where documents mix only over a single path. `TSNTM` then uses autoencoding variational Bayes to perform inference on the posterior probability of the distributions over paths and levels in the tree given the corpus.

`HETM` employs the same generative story and inference procedure. In the forward

pass, HETM first passes the bag-of-words representation of a document through a variational autoencoder to define a Gaussian distribution. HETM then samples from that Gaussian distribution and passes that sample through both a DRNN and RNN to draw a distribution over paths and levels down the tree. Formally, for the DRNN, HETM traverses the tree using breadth first search and applies the sigmoid activation function on the inner product of the Gaussian sample and the hidden state for each node. For the RNN, HETM traverses through each level of the tree and performs the same procedure. This is akin to sampling from two Beta distributions parametrized by the networks. HETM then refactors those samples into a distribution over paths and levels down the tree using the tree-based stick-breaking construction introduced by (Wang and Blei, 2009). The distributions over paths and levels are then used to calculate a distribution over nodes, or topics. HETM then uses a third DRNN to generate topic embeddings which are converted into probability distributions over the vocabulary by multiplying the hidden states of the DRNN with pretrained GloVe word embeddings (Pennington et al., 2014). Finally, the sample's distribution over topics and the topics' distributions over the vocabulary are used to reconstruct the original bag-of-words representation of the document. The final loss is then a weighted sum of the reconstruction loss, the KL loss from the variational autoencoder, and a regularizer on the topic embeddings to ensure topic diversity in low-dimensional space. This loss is then backpropagated through the network to train the HETM. Moreover, the tree structure is pruned and refined after concluding training for each epoch using a heuristic based on a user-defined proportion of words belonging to a single topic.

Unlike the TSNTM, HETM utilizes fixed and pretrained word embeddings to convert low-dimensional representations of topics into probability distributions over the vocabulary for document reconstruction. We also allow for trees with a depth greater than three. Moreover, we correct an error in the TSNTM's DRNN implementation. By activating the ancestral and fraternal states at each time step, we can vary the in-

put and output dimensions of the DRNN. Finally, we model the tree as its own data structure, simplifying storage of topic metadata. We perform Bayesian optimization to discover the best performing settings for our hyperparameters using the Ax library[3]. The complete set of hyperparameters, their default values, and descriptions are available in Table 5.

4.2 BERT-powered Stance Detector

We now turn to the stance detection section of the model: BERT-SD. BERT-SD is loosely based on the stance detection model introduced in Allaway and McKeown (2020). First, in cases where the target of stance detection is unknown, we generate a target. To do so, we pass the bag-of-words representation of a document through the forward pass of the HETM to retrieve the most likely topic as a probability distribution over the vocabulary. We then sample $p = 3$ words without replacement from the topic and combine those words into a string: our generated target.

BERT-SD takes two strings: a document and a target. These strings are embedded by BERT as a sentence pair. We then fit a two-layer feed-forward neural network on the classification token [CLS] embedding from the last hidden state of the model after performing inference on the sentence pair embedding. By using BERT, our model harnesses both contextualization in document and target embeddings as well as domain knowledge from pretraining on a massive corpus. This is especially helpful for detecting stance on novel generated targets or otherwise low-resource targets. We discuss the ethical considerations of using pretrained large-scale transformer-based language models in section (§7). The complete set of hyperparameters, their default values, and descriptions are available in Table 5.

[3]https://github.com/facebook/Ax

Hyperparameter	Description	Default	Model
var_hidden_size	variational autoencoder hidden size	32	HETM
rnn_hidden_size	RNNs and DRNNs input and hidden size	256	HETM
activation	the activation function used in the variational autoencoder	relu	HETM
max_depth	maximum permitted depth of the tree	3	HETM
max_sibling	maximum number of children for a single node	8	HETM
temperature	a value that encourages sparser probability distributions for topics deeper down the tree (Hinton et al., 2014)	10	HETM
enc_drop	the amount of dropout in the variational autoencoder	.5	HETM
clip	gradient clipping	1	HETM
optimizer	optimization algorithm used in training	adam	HETM
lr	learning rate for optimizer	.005	HETM
wdecay	weight decay for optimizer	1.2e-6	HETM
recon_param	coefficient for reconstruction loss	1	HETM
kl_param	coefficient for KL loss	1	HETM
reg_param	coefficient for regularization	1	HETM
prune_threshold	word diversity threshold at which tree removes node	k	HETM
refine_threshold	word diversity threshold at which tree adds child to node	k	HETM
early_stopping	boolean activated for early stopping	True	HETM, BERT-SD
patience	number of permitted epochs of insufficient validation loss decrease for early stopping	3	HETM, BERT-SD
min_delta_patience	minimum decrease in validation loss to not perform early stopping	0	HETM, BERT-SD
patience_every_n_epochs	epochs interval to check the validation loss for early stopping	1	HETM, BERT-SD
lr_scheduler	learning rate scheduler	True	HETM
min_lr	minimum learning rate for scheduler	1e-6	HETM
factor_lr	factor for learning rate scheduler	.5	HETM
p	number of words to sample from topic in target generation	3	BERT-SD

Table 5. Architecture hyperparameters

5 Experiments

5.1 Topic Modeling

We perform topic modeling experiments on the following models: LDA, hLDA, ETM, TSNTM, and HETM. We utilize the following datasets: SEMEVAL, DEBATEPEDIA, and VAST. A description of our baselines is available in section (§2.2) and a description of our datasets is available in section (§3.2). A description of our model, HETM, is available in section (§4.1).

Unlike standard topic modeling evaluations, our datasets include gold-labeled top-

ics: supertargets, subtargets, and the combination of those targets with each stance class. Therefore, we evaluate our models' ability to both model topics using targets as topics but also to detect stance using combination of those targets with each stance class as topics. As such, we do not evaluate our models on traditional topic modeling metrics such as topic interpretability (Isonuma et al., 2020; Dieng et al., 2020). Instead, we do not show the model some documents and some topics to test whether or not the model can place unseen documents with documents of the same topic together as well as documents of unseen topics together. To do so, we utilize two supervised coreference metrics, MUC and B^3, as well as two supervised clustering metrics, homogeneity and completeness. Moreover, we generalize our coreference metrics for settings such as `VAST` with multiple gold topics for a given documents.

Coreference is a concept from linguistics for differing expressions that refer to the same entity. For example, an antecedent and a proform may refer to the same subject: "Olivia likes to dance. Her preferred form is tap." In our example, the antecedent, Olivia, and proform, her, refer to the same person. We define a *link* as the entity itself, in our example, the person Olivia, and a *reference* as a linguistic representation of that entity, in our example, the terms "Olivia" and "her". We extend coreference to our setting by redefining *references* as documents and *links* as topics to evaluate topic models in settings where gold-labeled topics are known. For both metrics, we define a key entity set, or gold label set, G, where each element in the set is a list of documents corresponding to a known topic and we define response entity set, R, where each element in the set is a list of documents corresponding to a predicted topic. Moreover, we extend our metrics to handle settings where there are multiple possible arrangements of documents into topics because each document corresponds to multiple topics. In other words, settings with a set of N key entity sets, $\mathbb{G} = \{G\}_{i=1}^{N}$. MUC is a coreference metric introduced by Vilain et al. (1995) on links, or topics. For precision, MUC counts the number of "missing" documents

in the overlap of the gold topics compared to the predicted topics whereas for recall, MUC counts the number of "missing" documents in the overlap of the predicted topics compared to the gold topics. We define MUC Precision in Equation 2 and MUC Recall in Equation 3.

$$MUC_P = \frac{\sum_{r_j \in R} |r_j| - \sum_{g_i \in G} |g_i \cap r_j|}{\sum_{r_j \in R} |r_j| - 1} \tag{2}$$

$$MUC_R = \frac{\sum_{g_i \in G} |g_i| - \sum_{r_j \in R} |g_i \cap r_j|}{\sum_{G_i \in G} |g_i| - 1} \tag{3}$$

Moreover, we extend MUC Precision and Recall to settings with multiple key entity sets in Equations 4 and 5 respectively.

$$MUC_P = max_{G \in \mathbb{G}} \frac{\sum_{r_j \in R} |r_j| - \sum_{g_i \in G} |g_i \cap r_j|}{\sum_{r_j \in R} |r_j| - 1} \tag{4}$$

$$MUC_R = max_{G \in \mathbb{G}} \frac{\sum_{g_i \in G} |g_i| - \sum_{r_j \in R} |g_i \cap r_j|}{\sum_{G_i \in G} |g_i| - 1} \tag{5}$$

B^3 is a coreference metric introduced by Bagga and Baldwin (1998) on references, or documents, that counts difference in size of predicted and gold topic clusters for a given document. More specifically, B^3 Recall measures the extent to which predicted topics approximate gold topics as a fraction over gold topics while B^3 Precision measures the extent to which gold topics approximate predicted topics as a fraction over predicted topics.

We define B^3 Precision in Equation 6 and B^3 Recall in Equation 7.

$$B^3_P = \frac{\sum_{r_j \in R} \sum_{g_i \in G} \frac{|r_j \cap g_i|^2}{|r_j|}}{\sum_{r_j \in R} |r_j|} \tag{6}$$

$$B_R^3 = \frac{\sum_{r_j \in R} \sum_{g_i \in G} \frac{|r_j \cap g_i|^2}{|g_i|}}{\sum_{g_i \in G} |g_i|} \tag{7}$$

Moreover, we extend B^3 Precision and Recall to settings with multiple key entity sets in Equations 8 and 9 respectively.

$$B_P^3 = max_{G \in \mathbb{G}} \frac{\sum_{r_j \in R} \sum_{g_i \in G} \frac{|r_j \cap g_i|^2}{|r_j|}}{\sum_{r_j \in R} |r_j|} \tag{8}$$

$$B_R^3 = max_{G \in \mathbb{G}} \frac{\sum_{r_j \in R} \sum_{g_i \in G} \frac{|r_j \cap g_i|^2}{|g_i|}}{\sum_{g_i \in G} |g_i|} \tag{9}$$

We caution that our generalized metrics are not tractable. As such, we also introduce a greedy heuristic to compute these metrics. We iterate over the documents in the corpus, computing the highest scoring key entity set $G \in \mathbb{G}$ for the given document while preserving assignments for previously visited documents in the set. We modify the official CoNLL-2012 evaluation scripts to implement these two metrics (Pradhan et al., 2012).

Homogeneity and completeness are two popular supervised clustering metrics formalized in Rosenberg and Hirschberg (2007) and implemented using Scikit-Learn (Pedregosa et al., 2011). Clustering techniques are widely applied in natural language processing in settings where labeled data is not available; however, lack of a gold standard complicates interpretation of evaluation scores. Since our datasets do include labels for targets, we can harness clustering techniques to evaluate our topic modeling by considering documents referring to the same targets to be in the same cluster. For both metrics, we consider a corpus of N documents with a set of gold topic clusters G of size n, a set of predicted topic clusters R, and a contingency table $T = \{x_{ij}\}$ where x_{ij} is the number of documents that belong to gold topic cluster i and predicted topic cluster j. Homogeneity is a metric which is satisfied when all

predicted clusters each contain only documents that are members of the same gold cluster. In other words, homogeneity penalizes diversity in the predicted topic clusters with respect to the gold topic clusters. Formally, we define homogeneity h with Equation 10.

$$h = \begin{cases} 1 & \text{if } H(G,R) = 0 \\ 1 - \frac{H(G|R)}{H(G)} & \text{else} \end{cases} \tag{10}$$

where

$$H(G \mid R) = -\sum_{r=1}^{|R|}\sum_{g=1}^{|G|} \frac{a_{gr}}{N} \log \frac{a_{gr}}{\sum_{g=1}^{|G|} a_{gr}}$$

$$H(G) = -\sum_{g=1}^{|G|} \frac{\sum_{r=1}^{|R|} a_{gr}}{n} \log \frac{\sum_{r=1}^{|R|} a_{gr}}{n}$$

Completeness is a metric which is satisfied when all documents that are members of the same gold cluster are assigned to the same predicted cluster. In other words, completeness captures the extent to which the predicted topics capture all of the documents in a given gold topic. In this way, completeness is the complementary metric to homogeneity. Formally, we define completeness c with Equation 11.

$$c = \begin{cases} 1 & \text{if } H(R,G) = 0 \\ 1 - \frac{H(R|G)}{H(R)} & \text{else} \end{cases} \tag{11}$$

where

$$H(R \mid G) = -\sum_{g=1}^{|G|}\sum_{r=1}^{|R|} \frac{a_{gr}}{N} \log \frac{a_{gr}}{\sum_{r=1}^{|R|} a_{gr}}$$

$$H(R) = -\sum_{r=1}^{|R|} \frac{\sum_{g=1}^{|G|} a_{gr}}{n} \log \frac{\sum_{g=1}^{|G|} a_{gr}}{n}$$

In this paper, we report V-measure (Rosenberg and Hirschberg, 2007), the harmonic mean between homogeneity and completeness.

For SEMEVAL, we evaluate each model on both subtargets and subtarget-stance pairs. For DEBATEPEDIA and VAST, we evaluate each model on supertargets, subtargets, and subtarget-stance pairs. For VAST, we utilize our generalized coreference metrics for documents with mutliple supertargets, subtargets, and subtarget-stance pairs. Furthermore, we do not report V-measure scores for VAST because those metrics are not generalized to our multiple gold clustering solutions.

5.2 Stance Detection on Known Targets

We perform stance detection on known targets experiments on the following models: SVM-TFIDF, LR-TFIDF, BERT-SD. We utilize the following datasets: SEMEVAL, DEBATEPEDIA, and VAST. SVM-TFIDF is a baseline stance detection model that uses a support vector machine to perform classification. That model classifies the concatenated TFIDF representations of a document and target, restricted to 1000 features and fitted on the documents in the training set. We perform a grid search with the development set on the C, gamma, and kernel hyperparameters. LR-TFIDF is a baseline stance detection model that uses logistic regression to perform classification. That model classifies the same TFIDF representations as SVM-TFIDF, instead performing a grid search over the solvers of the regression. A description of our model, BERT-SD, is available in section (§4.2). Our evaluation technique for this task is simple: the macro-averaged F1 score of all classes for each document-subtarget example.

5.3 Stance Detection on Unknown Targets

We perform stance detection on unknown targets experiments on the following models: HETM + BERT-SD. We utilize the following datasets: SEMEVAL, DEBATEPEDIA, and VAST. Unlike stance detection on known targets, we are not able to evaluate the performance of our model simply on the macro-average F1 of all classes for each document-subtarget example. This is because we do not know to what extent the

generated target approximates the gold target in framing or meaning. However, we do not have any other mechanism to automatically determine the gold stance on a generated target other than harness the information encoded by the gold stance on a known target. As such, we compute a *sample-weighted* F1 score of all classes. We weight each example by the extent to which the generated target is semantically similar to the known target. We assume that generated targets that are similar to known targets ought to have similar stances whereas generated targets that are dissimilar to known targets ought to have dissimilar stances. Therefore, we place more weight on examples where we expect the predicted stance to match the gold stance and less weight on examples where we expect predicted stance to differ from the gold stance. We compute a weight $\hat{\theta}$ by computing cosine similarity θ on the classification token `[CLS]` embeddings from the last hidden state of `BERT` after performing inference on the the generated and known target separately. We use $\hat{\theta} = \frac{\theta+1}{2}$ to transform cosine similarity into $[0, 1]$ for use as our weight. Moreover, we only generate a single target per document but compare that pair to each gold target available for that document.

6 Results and Discussion

6.1 Topic Modeling

In this section, we discuss the results of our topic modeling experiments. While some patterns we note may be caused by idiosyncrasies in our models, hyperparameters, datasets, or even in the metrics themselves, we note some signals in the noise. Most striking, on our low resource dataset, `SEMEVAL`, hierarchical and embedding topic models, including `HETM`, outperform other models. Moreover, on our hierarchical dataset, `DEBATEPEDIA`, hierarchical models outperform other models. This supports our claim that hierarchical and embedding topic models are better suited for modeling low resource settings and that hierarchical models are better suited for modeling

	MUC					
	DEBATEPEDIA			VAST		
Model	Train	Dev	Test	Train	Dev	Test
LDA	.853	.703	.713	.330	.343	.347
hLDA	.907	.817	.842	.307	.394	.299
ETM	.807	.606	.620	.198	.245	.242
TSNTM	.995	.984	.988	.155	.210	.190
HETM	.984	.957	.961	.157	.213	.182
	B^3					
	DEBATEPEDIA			VAST		
Model	Train	Dev	Test	Train	Dev	Test
LDA	.099	.121	.105	.171	.215	.180
hLDA	.279	.317	.258	.284	.370	.272
ETM	.051	.082	.069	.097	.155	.131
TSNTM	.232	.262	.184	.155	.210	.190
HETM	.227	.259	.183	.156	.213	.181
	V-measure					
	DEBATEPEDIA			VAST		
Model	Train	Dev	Test	Train	Dev	Test
LDA	.134	.194	.167	-	-	-
hLDA	.193	.253	.203	-	-	-
ETM	.055	.144	.121	-	-	-
TSNTM	.000	.000	.000	-	-	-
HETM	.007	.019	.017	-	-	-

Table 6. Topic modeling metrics for supertargets, F1 scoring

hierarchical settings. We argue that such models are better able to learn and represent topics with limited data by harnessing dimensionality reduction, domain knowledge from pretrained embeddings as well as modeling relationships between topics. This holds true both for target and target-stance settings, implying that the same methods we use to model topics may also be employed in detecting stance. A bridge between the two tasks would open new horizons for stance detection in low resource and unknown settings.

In modeling supertargets, we uncover complex and contradictory results among the models. For example, on *MUC* all models performed better on DEBATEPEDIA whereas for B^3, all models except HETM performed better on VAST. This difference

	MUC								
	SEMEVAL			DEBATEPEDIA			VAST		
Model	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
LDA	.991	.991	.987	.283	.190	.154	.404	.242	.275
hLDA	.969	.892	.954	.663	.489	.510	.201	.429	.292
ETM	.991	.965	.987	.037	.011	.014	.014	.007	.000
TSNTM	.993	.973	.990	.822	.627	.678	.023	.135	.093
HETM	.993	.976	.991	.885	.751	.798	.025	.149	.105
	B^3								
	SEMEVAL			DEBATEPEDIA			VAST		
Model	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
LDA	.235	.235	.223	.133	.304	.207	.499	.508	.556
hLDA	.253	.262	.266	.235	.343	.260	.202	.416	.295
ETM	.206	.208	.210	.073	.251	.179	.885	.614	.744
TSNTM	.288	.281	.295	.012	.017	.016	.023	.132	.093
HETM	.328	.333	.333	.012	.019	.017	.025	.148	.105
	V-measure								
	SEMEVAL			DEBATEPEDIA			VAST		
Model	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
LDA	.037	.034	.014	.526	.699	.614	-	-	-
hLDA	.078	.104	.065	.550	.650	.562	-	-	-
ETM	.003	.005	.004	.503	.714	.647	-	-	-
TSNTM	.002	.005	.003	.033	.102	.079	-	-	-
HETM	.002	.020	.006	.031	.072	.056	-	-	-

Table 7. Topic modeling metrics for subtargets, F1 scoring

was most pronounced for *MUC* and for neural methods. Neural methods typically outperformed traditional methods; however, for V-measure, every traditional method outperformed every neural method. On DEBATEPEDIA, hierarchical models outperformed traditional models, whereas on VAST, this property only held for the B^3. HETM performs well on *MUC* and B^3 metrics without a discernible difference in relative performance between the two datasets; however, our model performs poorly on V-measure. Second, in modeling subtargets, we find more promising results. On SEMEVAL, neural methods perform best while on DEBATEPEDIA, hierarchical models are the most successful. However, we note that neural and hierarchical models generally perform poorly on V-measure while traditional models outperform neural models

	MUC								
	SEMEVAL			DEBATEPEDIA			VAST		
Model	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
LDA	.913	.913	.881	.201	.110	.107	.274	.271	.266
hLDA	.802	.578	.776	.478	.329	.308	.228	.381	.308
ETM	.912	.688	.875	.010	.003	.006	.000	.007	.007
TSNTM	.978	.914	.969	.663	.342	.500	.016	.103	.067
HETM	.975	.914	.966	.723	.479	.565	.014	.101	.071
	B^3								
	SEMEVAL			DEBATEPEDIA			VAST		
Model	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
LDA	.103	.103	.100	.156	.375	.281	.314	.414	.407
hLDA	.124	.146	.136	.268	.473	.359	.230	.438	.362
ETM	.081	.095	.084	.137	.421	.318	.940	.665	.796
TSNTM	.134	.138	.159	.007	.011	.012	.016	.101	.067
HETM	.149	.157	.152	.007	.013	.011	.014	.100	.071
	V-measure								
	SEMEVAL			DEBATEPEDIA			VAST		
Model	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
LDA	.053	.052	.049	.594	.752	.685	-	-	-
hLDA	.154	.250	.141	.721	.828	.768	-	-	-
ETM	.164	.063	.023	.654	.829	.775	-	-	-
TSNTM	.004	.013	.044	.057	.163	.158	-	-	-
HETM	.005	.029	.010	.061	.145	.122	-	-	-

Table 8. Topic modeling metrics for subtargets-stance, F1 scoring

on DEBATEPEDIA using B^3. Finally, on VAST, flat and traditional models outperform hierarchical and embedding models; however, we recognize that the training performance on hierarchical and embedding models is also poor. In fact, performance on the test set improved for hierarchical and embedding models while it decreased for other models.

Our modeling of subtarget-stance pairs confirms many of the patterns we recognized in simple target modeling. However, unlike typical topic modeling settings, this setting further encodes information about the stances taken in documents. As such, these results ought to be carefully considered in harnessing topic modeling for stance detection and related sentiment classification tasks. We note that an important way

that stances may be explicitly detected using topic modeling is through the use of framing or agenda setting whereby authors use key words or refer to key aspects of a debate in service of making their point clear. On both DEBATEPEDIA and SEMEVAL, hierarchical embedding and hierarchical models perform well on MUC while hierarchical embedding models perform poorly on B^3 and V-measure. However, such poor performance is reflected in both training and test test. As such, we caution that such measures must be analyzed in the broader context of a model's training process. A model that performs poorly on a training set ought to be expected to perform poorly on a test set. Finally, on VAST, traditional models perform best while hierarchical embedding models perform worst. We also note the strong performance of ETM; however, it does benefit from successful learning on the training set over other models. We recognize that many of the results we see may be influenced by default hyperparameters of our baselines or by insufficient hyperparameter searches in our models. Moreover, some metrics, such as V-measure, may not be suited for studying the performance of a topic model in a supervised setting. Additional steps must be taken to develop metrics and datasets that are better equipped to measure and study our task — hierarchical topic modeling for low-resource settings.

6.2 Stance Detection

	SEMEVAL								
	F1			Pr			Re		
Model	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
SVM-TFIDF	.996	.996	.602	.995	.996	.667	.996	.997	.583
LR-TFIDF	.758	.741	.527	.844	.818	.549	.727	.712	.519
BERT-SD	.748	.595	.532	.746	.593	.533	.768	.611	.558
HETM+BERT-SD	.324	.303	.325	.331	.314	.337	.332	.310	.338

Table 9. Stance detection metrics for SEMEVAL, macro-averaged
All models perform stance detection on known targets,
except HETM+BERT-SD which performs stance detection on unknown targets

	DEBATEPEDIA								
	F1			Pr			Re		
Model	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
SVM-TFIDF	.995	.999	.686	.996	.999	.701	.995	.999	.675
LR-TFIDF	.920	.911	.669	.920	.911	.676	.919	.911	.663
BERT-SD	.628	.462	.427	.628	.457	.426	.637	.473	.438
HETM+BERT-SD	.269	.240	.230	.268	.247	.237	.262	.245	.234

Table 10. Stance detection metrics for DEBATEPEDIA, macro-averaged
All models perform stance detection on known targets,
except HETM+BERT-SD which performs stance detection on unknown targets

	VAST								
	F1			Pr			Re		
Model	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
SVM-TFIDF	.712	.714	.620	.737	.718	.637	.711	.717	.624
LR-TFIDF	.850	.869	.734	.854	.873	.738	.848	.868	.738
BERT-SD	.753	.595	.585	.754	.600	.590	.757	.599	.589
HETM+BERT-SD	.500	.499	.498	.504	.501	.497	.498	.501	.497

Table 11. Stance detection metrics for VAST, macro-averaged
All models perform stance detection on known targets,
except HETM+BERT-SD which performs stance detection on unknown targets

In this section, we discuss the results of our stance detection experiments. We note that we perform two variations of stance detection. On three models (SVM-TFIDF, LR-TFIDF, BERT-SD), we perform the traditional stance detection task whereby a model receives as input a document and target, classifying stance on the pair. On one model (HETM+BERT-SD), we perform a novel stance detection task whereby a model generates a target from a document using a topic model and then classifies stance on the document-generated target pair. Since we do not have gold labeled stances on generated topics, we instead weight the prediction by the cosine similarity of the generated and known targets when calculating loss on that known target's gold stance.

We generally find strong performance by both the baseline models and BERT-SD, with HETM+BERT-SD performing slightly worse on the novel task. Moreover, we find

that baselines typically outperform BERT-SD, except on SEMEVAL. Moreover, we find that both BERT-SD and HETM+BERT-SD perform best on VAST. We expect model like BERT-SD that are pretrained on a massive corpus to outperform naïve models on both SEMEVAL, our lowest resource dataset and VAST, which includes many low and zero resource targets. This is because a pretrained model encodes domain knowledge that may then be used in service of performing a task such as stance detection that relies on a model's ability to understand the world. While our pretrained models performed best on VAST, they performed worst on DEBATEPEDIA, which was also the dataset where there was the largest gap in performance between the pretrained models and the baselines. In general, models performed relatively consistently; however, there was a 20 point increase for LR-TFIDF and HETM+BERT-SD from SEMEVAL to VAST. This demonstrates that these models are better able to handle large datasets with low resource targets than small datasets with high resource targets.

7 Ethical Considerations

In this section, we discuss the ethical considerations of work. First, we discuss our use of pretrained language models in our architecture, namely a large-scale transformer-based language models, BERT (Devlin et al., 2018), and a pretrained word embedding model, GLoVe (Pennington et al., 2014). Pretrained language models are known to encode and exhibit social biases due to their training corpora consisting of extremely large and uncurated scrapes of the Internet (Caliskan et al., 2017; Bender et al., 2021). While some work has been done to "debias" these models, these methodologies have been shown to be inherently insufficient in mitigating these biases (Bolukbasi et al., 2016; Gonen and Goldberg, 2019). While pretrained models are a powerful tool that lay claim to many states-of-the-art in natural language processing, their use must always be cautioned as a potential vector for the introduction of social bias into

downstream tasks. This is especially true for models that are used in real world decision-making.

Second, we discuss the ethical considerations of our task: stance detection. Work on detecting social biases learned by models during training on sentiment analysis tasks, such as stance detection, is limited to emotion detection (Kiritchenko and Mohammad, 2018; Câmara et al., 2022). An issue with studying algorithmic fairness for stance detection is that individuals who belong to the same disadvantaged class may hold similar views for a myriad of social, political, or cultural reasons. As such, there is a risk of a model predicting certain stances for a set of authors not based on their position but on identifying markers of identity. This concern is especially pronounced for groups such as Black Americans who may communicate in African American English. As such, a model may conflate uses of African American English with stances that are popular, but not universal, within the Black community.

We also point to potential limitations in the study of Black communities by natural language processing researchers. In this paper, we collect several prospective stance datasets that reflect the Black community. However, we are limited in our ability to study these datasets both due to technical reasons discussed in section (§3.1) and also because we are not members of this community. As such, we cannot successfully or ethically annotate a dataset both because we do not know exactly which topics are of most concern to members of the Black community but also because we are not speakers of African American English. While we do have some domain knowledge of American life and language that may transfer to this setting, it is important to recognize that the experiences of members of the Black community are unique in character and must be carefully studied. Moreover, members of any group ought to share in the responsibilities and rewards of research on their group. Diversity within the computer science and natural language processing community is critical to both the advancement of the discipline and of marginalized groups.

8 Future Work

In this section, we discuss future directions for our work. First, we are interested in extensions to the stance detection component of our work. First, we are interested in the concept of *framing* from the social sciences, which posits that issues may be understood through different lenses which in turn drives how individuals form and express their attitudes (Chong and Druckman, 2007). Existing work in stance detection, with the exception of Allaway and McKeown (2020), neglects this concept. We believe that this concept is paramount in understanding how authors express stance and understand the world. Second, we are interested in developing stance detection datasets and models that are better suited to model stance and targets in both space and time. For example, we are interested in developing time-series stance detection datasets and models to understand how attitudes change over time and in response to events. We also hope to continue our work in hierarchical stance detection to better understand the relationships between targets and how those relationships affect stance.

Second, we are interested in extensions to the topic modeling component of our work. We hope to develop new methods in reconstruction loss for unsupervised learning. Existing methods in embedding topic modeling reconstruct bag-of-words representations of documents, much like traditional methods. We hope to instead reconstruct low-dimensional representations of documents. We believe such a technology will aid in learning more robust topics and more robust low-dimensional semantic spaces that may be used in the service of topic modeling and stance detection.

Third, we are interested in bringing stance detection and topic modeling closer together. We aim to do this by developing a single-model architecture for stance detection on unknown targets where such targets are discovered using topic modeling. We hope to tools such as contrastive learning to do this. Moreover, we are

interested in continuing our experiments in detecting stance using topic models. We believe that we can use the semantic representations learned by embedding models in tandem with methods from topic modeling to better perform this task. Finally, we are interested in extending our work into low-resource settings. For example, we hope to develop completed stance detection datasets using the preliminary datasets we collected to study the task in Black community and African American English setting. We are also interested in extending our work to study other minority groups, communities, dialects, and languages. In doing so, we plan to center considerations of algorithmic fairness and interpretability in our work to ensure that the methods we develop successfully and ethically serve researchers across disciplines and the public.

9 Conclusion

In this paper, we introduce a novel task, stance detection on unknown targets, which build on two existing tasks, stance detection and topic modeling. We discuss these three tasks and their limitations in studying low-resource settings where the topics of interest are unknown and related. For example, we discuss existing natural language processing resources for the study of African American English and Black communities as well as introduce sample datasets to that end. We also introduce two models, the Hierarchical Embedding Topic Model (`HETM`) and the BERT-powered Stance Detector (`BERT-SD`) in service of performing these three tasks. Our models perform well against popular baselines in stance detection and topic modeling. We also experiment with our complete architecture on our novel task. Finally, we consider the ethical implications of and potential future directions for our work.

References

Allaway, Emily and Kathleen McKeown, "Zero-shot stance detection: A dataset and model using generalized topic representations," *arXiv preprint arXiv:2010.03640*, 2020.

Alvarez-Melis, David and Tommi S Jaakkola, "Tree-structured decoding with doubly-recurrent neural networks," 2016.

Augenstein, Isabelle, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva, "Stance detection with bidirectional conditional encoding," *arXiv preprint arXiv:1606.05464*, 2016.

Bagga, Amit and Breck Baldwin, "Algorithms for scoring coreference chains," in "The first international conference on language resources and evaluation workshop on linguistics coreference," Vol. 1 Citeseer 1998, pp. 563–566.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in "Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency" FAccT '21 Association for Computing Machinery New York, NY, USA 2021, p. 610–623.

Blei, David M, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, 2003, *3* (Jan), 993–1022.

Bois, John W Du, "The stance triangle," *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 2007, *164* (3), 139–182.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, 2016, *29*.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, 2017, *356* (6334), 183–186.

Câmara, António, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel, "Mapping the Multilingual Margins: Intersectional Biases of Sentiment Analysis Systems in English, Spanish, and Arabic," *arXiv preprint arXiv:2204.03558*, 2022.

Chen, Emily, Ashok Deb, and Emilio Ferrara, "# Election2020: the first public Twitter dataset on the 2020 US Presidential election," *Journal of Computational Social Science*, 2021, pp. 1–18.

_ , **Kristina Lerman, and Emilio Ferrara**, "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set," *JMIR Public Health and Surveillance*, 2020, *6* (2), e19273.

Chong, Dennis and James N Druckman, "Framing theory," *Annu. Rev. Polit. Sci.*, 2007, *10*, 103–126.

DeVerna, Matthew R, Francesco Pierri, Bao Tran Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Christopher Torres-Lugo, Kai-Cheng Yang, Filippo Menczer, and John Bryden, "CoVaxxy: A collection of English-language Twitter posts about COVID-19 vaccines," in "Proceedings of the AAAI international conference on web and social media (ICWSM)" 2021.

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

Dieng, Adji B, Francisco JR Ruiz, and David M Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, 2020, *8*, 439–453.

Erikson, Robert S and Kent L Tedin, *American public opinion: Its origins, content and impact*, Routledge, 2015.

Fishbein, Martin and Icek Ajzen, "Attitudes and opinions," *Annual review of psychology*, 1972, *23* (1), 487–544.

Giorgi, Salvatore, Sharath Chandra Guntuku, Muhammad Rahman, McKenzie Himelein-Wachowiak, Amy Kwarteng, and Brenda Curtis, "Twitter corpus of the# blacklivesmatter movement and counter protests: 2013 to 2020," *arXiv preprint arXiv:2009.00596*, 2020.

Gonen, Hila and Yoav Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them," *arXiv preprint arXiv:1903.03862*, 2019.

Gottopati, Swapna, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah Smith, "Learning topics and positions from debatepedia," in "in" ACL 2013.

Griffiths, Thomas, Michael Jordan, Joshua Tenenbaum, and David Blei, "Hierarchical topic models and the nested Chinese restaurant process," *Advances in neural information processing systems*, 2003, *16*.

Habernal, Ivan, Henning Wachsmuth, Iryna Gurevych, and Benno Stein, "The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants," in "Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, Volume 1 (Long Papers)" Association for Computational Linguistics New Orleans, Louisiana June 2018, pp. 1930–1940.

Hasan, Kazi Saidul and Vincent Ng, "Stance classification of ideological debates: Data, models, features, and constraints," in "Proceedings of the sixth international joint conference on natural language processing" 2013, pp. 1348–1356.

Hinton, Geoffrey, Oriol Vinyals, and Jeffrey Dean, "Distilling the Knowledge in a Neural Network. NIPS 2014 Deep Learning Workshop," *arXiv preprint arXiv:1503.02531*, 2014.

Isonuma, Masaru, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata, "Tree-structured neural topic model," in "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics" 2020, pp. 800–806.

Kiritchenko, Svetlana and Saif M. Mohammad, "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems," *CoRR*, 2018, *abs/1805.04508.*

Küçük, Dilek and Fazli Can, "Stance detection: A survey," *ACM Computing Surveys (CSUR)*, 2020, *53* (1), 1–37.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin, "Advances in Pre-Training Distributed Word Representations," in "Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)" 2018.

Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in "Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)" 2016, pp. 31–41.

Muric, Goran, Yusong Wu, Emilio Ferrara et al., "COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies," *JMIR public health and surveillance*, 2021, *7* (11), e30642.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in "Advances in Neural Information Processing Systems 32," Curran Associates, Inc., 2019, pp. 8024–8035.

Pedregosa, F., 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. Duch-

esnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2011, *12*, 2825–2830.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in "Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)" 2014, pp. 1532–1543.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang, "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes," in "Joint Conference on EMNLP and CoNLL - Shared Task" Association for Computational Linguistics Jeju Island, Korea July 2012, pp. 1–40.

Rosenberg, Andrew and Julia Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in "Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)" 2007, pp. 410–420.

Sobhani, Parinaz, Diana Inkpen, and Xiaodan Zhu, "A dataset for multi-target stance detection," in "Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers" 2017, pp. 551–557.

Taulé, Mariona, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, Viviana Patti et al., "Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017," in "2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017," Vol. 1881 CEUR-WS 2017, pp. 157–177.

Thomas, Matt, Bo Pang, and Lillian Lee, "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts," *arXiv preprint cs/0607062*, 2006.

Vamvas, Jannis and Rico Sennrich, "X-stance: A multilingual multi-target dataset for stance detection," *arXiv preprint arXiv:2003.08385*, 2020.

Vayansky, Ike and Sathish AP Kumar, "A review of topic modeling methods," *Information Systems*, 2020, *94*, 101582.

Vilain, Marc, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman, "A model-theoretic coreference scoring scheme," in "Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995" 1995.

Wang, Chong and David Blei, "Variational inference for the nested Chinese restaurant process," *Advances in Neural Information Processing Systems*, 2009, *22*.

Xu, Chang, Cecile Paris, Surya Nepal, and Ross Sparks, "Cross-target stance classification with self-attention networks," *arXiv preprint arXiv:1805.06593*, 2018.