Tweet Classification without the Tweet: An Empirical Examination of User versus Document Attributes

Veronica E. Lynn[1], **Salvatore Giorgi**[2],
Niranjan Balasubramanian[1] and **H. Andrew Schwartz**[1]
[1]Stony Brook University, [2]University of Pennsylvania
{velynn, niranjan, has}@cs.stonybrook.edu, sgiorgi@sas.upenn.edu

Abstract

NLP naturally puts a primary focus on leveraging document language, occasionally considering user attributes as supplemental. However, as we tackle more social scientific tasks, it is possible user attributes might be of primary importance and the document supplemental. Here, we systematically investigate the predictive power of user-level features alone versus document-level features for *document*-level tasks. We first show user attributes can sometimes carry more task-related information than the document itself. For example, a tweet-level stance detection model using only 13 user-level attributes (i.e. features that did not depend on the specific tweet) was able to obtain a higher F1 than the top-performing SemEval participant. We then consider multiple tasks and a wider range of user attributes, showing the performance of strong document-only models can often be improved (as in stance, sentiment, and sarcasm) with user attributes, particularly benefiting tasks with stable "trait-like" outcomes (e.g. stance) most relative to frequently changing "state-like" outcomes (e.g. sentiment). These results not only support the growing work on integrating user factors into predictive systems, but that some of our NLP tasks might be better cast primarily as user-level (or human) tasks.

1 Introduction

Natural language processing is increasingly tackling new tasks over microblogs and social media, such as stance detection, sarcasm detection, and variations of sentiment analysis. Building on techniques used for traditional NLP, it is natural to attempt such tasks with inputs based solely on the content of the document (e.g. tweet) in question. We present an empirical argument for why this text-only scope may be a limiting view which inflates the value of document-only solutions.

Our work aims to answer the following questions: 1) What and how much information do user attributes alone carry for different social media tasks, particularly for predictive tasks that are more about the user than the document (e.g. stance)? 2) When are user attributes useful and what do language features contribute in these cases? While there are multiple works that show that adding user attributes is useful for different prediction tasks (Hovy, 2015; Zamani and Schwartz, 2017; Lynn et al., 2017), there is no single systematic study that answers these questions.

To this end we conduct a systematic evaluation of user attribute-only models on multiple tasks including stance detection, sarcasm detection, sentiment analysis, and prepositional phrase attachment. We evaluate the impact of user attribute-only models through a range of features derived from publicly available information about the users including: written profile bio, inferred demographics and personality, self-reported location, profile picture, who one follows in a social network, and a background of users' past language. The evaluations show that user attributes can have a large impact and, depending on the nature of the task, even outperform document-only features — *inference on a document without even looking at its contents!*

We conduct further evaluations comparing document contributions to an inference task relative to user-level features. Recent research has explored *how user-level attributes add value* ***on top of*** *document-level language* (Volkova et al., 2013; Hovy, 2015; Lynn et al., 2017; Zamani et al., 2018). Instead we quantify how well user attributes **alone** can predict and then what document-level language can uniquely add, identifying cases where the document is essential.

Contributions. Our specific contributions are three-fold: **(1)** We show that the stance of a

Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science, pages 18–28
Minneapolis, Minnesota, June 6, 2019. ©2019 Association for Computational Linguistics

tweet can be predicted with state-of-the-art F1 scores (better than all participant systems of the SemEval-2016 stance task) without even looking at the given tweet, suggesting such tasks might be better case as user-level (we outperform tweet-specific models that use thousands of features or complex neural networks using only 13 easily-derived features). **(2)** We put forth a theory that tasks which capture more "trait-like" human attributes (those that are stable over time, e.g. stance) benefit more from user-level information as compared to "state-like" attributes (frequently changing, e.g. sentiment). We evaluate this theory by looking at the role of user attributes across different predictive tasks. **(3)** We provide a set of considerations and metrics, for task participants and designers alike, for the inclusion of user information within new social science-related tasks.

2 Background

Recent work has shown that considering language within the context of user attributes can improve classification accuracy (Volkova et al., 2013; Bamman et al., 2014; Yang and Eisenstein, 2015; Hovy, 2015; Kulkarni et al., 2016; Lynn et al., 2017). Other work has used network or other meta data, such as in Bamman and Smith (2015); Johnson and Goldwasser (2016); Joseph et al. (2017); Khattri et al. (2015). In a sense these trail-blazing works might be viewed as case studies on user attributes — identifying particular pieces of information for particular tasks where user information has lead to an advantage. We believe this is the first systematic study on the extent to which tasks are more easily achieved with user information or by combining user attributes with document language. In addition, prior work has explored what user attributes add on top of language, whereas we focus primarily on user attributes, with the contributions from document-level features being secondary.

Models designed specifically to put language within the context of human factors, such as demographics or location, have led to improvements on a variety of NLP tasks. For example, Hovy (2015) improved on three types of text classification tasks by learning age- and gender-specific word embeddings. Similarly, Yang and Eisenstein (2015) found that sentiment analysis benefited from learning community-specific embeddings from social networks. Lynn et al. (2017) proposed a method to adapt language to user factors by composing the factors with language features in a domain adaptation-like formulation, demonstrating improvements on multiple tasks; this technique was expanded upon by Zamani et al. (2018). Still, even simple methods for incorporating these factors provide predictive power and should not be overlooked; our paper examines this in-depth.

Some work in stance detection has focused on document context and discourse structure (Walker et al., 2012a,b; Sridhar et al., 2015), though user attributes have been considered as well. When predicting stance for debates, Thomas et al. (2006) and Hasan and Ng (2013) benefited from enforcing the constraint that multiple statements from the same person should receive the same predicted stance, making the assumption that stance is unlikely to change over the course of a single conversation. Johnson and Goldwasser (2016), who predict the stance of Twitter *users* as opposed to individual tweets, consider both temporal activity and political party affiliation in their models. Chen and Ku (2016) learned user embeddings for stance detection and found that the inclusion of such embeddings significantly improved model performance. Going in a somewhat different direction, Joseph et al. (2017) found that the amount and type of user attributes, such as political party affiliation or Twitter profile description, provided to annotators of a stance detection dataset significantly impacted annotation quality, suggesting that considering user attributes is important not just during classification but also during dataset creation.

User attributes and other contextual information has proven useful beyond stance detection. Bamman and Smith (2015) extensively evaluate the effects of extralinguistic information, including *author*, *audience*, and *environment* features, in the context of sarcasm detection. They observe an almost 10 point increase in performance when adding extralinguistic features to the text-only model and find such features perform well even without the textual features. Although their work is similar to ours, we explore more tasks and a different set of extralinguistic features, including inferred factors; we see our work as complementary to — and expanding on — theirs.

Amir et al. (2016) outperformed Bamman and Smith (2015) on the same dataset by incorporating user embeddings, learned from users' past tweets, into a deep sarcasm model. Khattri et al. (2015) use past tweets to improve sarcasm detection by comparing the sentiment expressed towards an entity in the target tweet to that expressed in historical tweets. Martin et al. (2016) found that, when predicting retweet count, a user's past success (measured as the average number of retweets received for other tweets in the past) was nearly as predictive as a model using all features they tried, including those drawn from the tweet itself. Jurgens et al. (2017) find that they are able to accurately predict the attributes of a user based on communications targeted at them (as opposed to written by them), emphasizing that a person's social network is itself an important source of user-level information. Finally, Hovy and Fornaciari (2018) demonstrate that user attributes can be used to improve the quality of author embeddings via retrofitting.

3 Prediction Models

This paper seeks to systematically and empirically understand the role of user attributes within the context of social media tasks. To that end, we consider a variety of user-level features and evaluate their importance for four tweet-level prediction tasks.

3.1 Tasks

The following section provides details for the systems and datasets used for analysis. Development sets were used for hyperparameter tuning. Statistics for each task are given in Table 1.

Stance. For stance detection we use the SemEval-2016 dataset (Mohammad et al., 2016), which contains tweets annotated as being *in favor of*, *against*, or *neutral toward* one of five targets: atheism, climate change as a real concern, feminism, Hillary Clinton, and legalization of abortion. Note that *neutral* does not indicate "neither for nor against", but rather not enough information to say either way (for example, "I know who I'm voting for!" would be *neutral* towards Hillary Clinton). Similar to the top baseline system in this task, we train a logistic regression classifier on character n-grams of size two to five and word n-grams of size one to three. We preserve the train/test split of the original dataset. For evaluation purposes, we obtain the predictions from the top participating system, MITRE (Zarrella and Marsh, 2016), and subset them to our test set.

Sarcasm. Sarcasm detection replicates the work of Bamman and Smith (2015) by using the tweet features described in the paper (e.g n-grams, sentiment scores, Brown clusters) and evaluating on their dataset using a logistic regression classifier via ten-fold cross validation. The folds are split such that no user appears in both the training and testing sets. Bamman and Smith (2015)'s dataset was constructed by sampling tweets that did or did not contain hashtags indicating sarcasm (e.g. *I love when it snows #sarcastic*); these hashtags were removed during preprocessing.

Sentiment. Message-level sentiment annotations indicating *positive*, *negative*, and *neutral* are available from the SemEval-2013 dataset (Nakov et al., 2013). We mostly replicate the top-performing system on this task (Mohammad et al., 2013) by training a linear SVM on character n-grams, word n-grams, and features from multiple sentiment and emotion lexicons (Hu and Liu, 2004; Wilson et al., 2005; Mohammad and Turney, 2010, 2013; Mohammad et al., 2013; Kiritchenko et al., 2014). The train/test split of the original dataset was used for evaluation.

PP-Attachment. A prepositional-phrase attachment dataset for Twitter was constructed by combining annotated data from Tweebank (Kong et al., 2014) and Lynn et al. (2017). Candidate heads are ranked using an SVM-Rank (Joachims, 2006) model trained on n-gram, WordNet, and Treebank features similar to those used in Belinkov et al. (2014). Cross validation is used for evaluation.

Task	Tweets	Users	Instances
Stance	3021	2349	3021
Sarcasm	17084	10966	17084
Sentiment	10339	9917	10339
PP-Attachment	1319	1319	2365

Table 1: Number of tweets, users, and instances represented in each task.

3.2 User Attribute Features

Each user's name, location, description, and picture were extracted from their Twitter profile. We also collected up to 200 of their tweets, excluding retweets and those included in the task data. Finally, we collected a list of every account that each user follows. Features were derived from this data as described below. We excluded tweets for which no user information was available; as a result, the test and training datasets were typically smaller than the originals[1].

One concern with using predicted user attributes such as age, gender, or personality is that they are prone to noise. However, one can look at it as a way of reducing large quantities of text to a single feature that happens to correlate well with some external quantity. Because we were interested in what a person's language says about themselves, any discrepancies between a user's predicted and actual attribute may provide additional predictive power: a 50-year-old whose writing style is more typical of a 20-year-old is likely better represented using their predicted age (20) than their actual age (50).

Demographics & Personality. Real-valued estimates of these attributes were obtained by applying pre-existing predictive lexica to each user's past tweets. Age and gender were obtained from the models of Sap et al. (2014). For personality, we used Park et al. (2015) to predict each of the Big Five traits: openness, conscientiousness, extroversion, agreeableness, and neuroticism.

Political Ideology. Using the dataset from Preoţiuc-Pietro et al. (2017), we train a ridge regression model on topic and n-gram features to predict real-valued political ideology scores between 1 (*very conservative*) and 7 (*very liberal*) from each user's tweets. This model achieved a Pearson r = .374 through cross validation of the training data.

User Embeddings. Five-dimensional latent factors were derived from each user's prior tweets using the generative factor analysis approach proposed by Kulkarni et al. (2017). Factors obtained using this method have been shown to correlate with outcomes such as income and IQ.

Profile Name. We used the Demographer package (Knowles et al., 2016) to predict gender from the profile name. We also used NamePrism (Ye et al., 2017) to predict scores for six ethnicities and thirty-nine nationalities.

Profile Description & Location. Character 2- to 5-grams and word 1- to 3-grams were extracted from the users' description and location fields.

Profile Picture. Borrowing from a popular method in transfer learning, we used a pre-trained image classification model, Inception-v3 (Szegedy et al., 2016), to obtain 2048-dimensional embeddings from the next-to-last layer of the model.

Followees. For each task, we identified the top 5000 Twitter accounts that were followed by the users in our dataset. Each of the 5000 accounts corresponded to a binary feature indicating if the user followed that account or not. We chose this representation for simplicity, though alternative methods such as network embeddings (e.g. Yang and Eisenstein (2015)) may be used instead.

4 Stance of a Tweet without the Tweet

We look first at the task of stance detection, as stance is typically seen more as a *trait* (attributable to a user) than a *state* (attributable to a point in time, such as a single message).

Table 2 compares stance prediction results for models trained only on tweet features to those trained only on user attribute features. Here, we only directly consider the *favor* and *against* classes so as to be consistent with the SemEval competition, which used an F1 measure that is an unweighted average of just these two classes. Note that `Inferred Factors` is a combination of `Demographics`, `Personality`, `Political Ideology`, and `User Embeddings`.

Stance without tweet is better than tweet only: Two of the user attribute types, `Followee` and `Inferred Factors`, perform better than the best tweet-based system that participated in SemEval-2016. `Location` also performs better than the most frequent class baseline. As we show next, if we consider the performance on the *neutral* class, we find that user attributes can do even better. We expect user attributes to carry some stance related information but it is surprising that they can compete with or outperform state-of-the-art models despite using a simpler model and/or fewer features.

Table 3 shows results when considering the full three-way classification task, where we evaluate performance on the *favor*, *against*, and *neutral* classes by taking the weighted average of F1 in all three classes. The table compares against the most frequent class (MFC) baseline to illustrate how much stance-related information is contained in each of the user attributes.

User attributes carry useful information for all stance prediction tasks as shown in Table 3. `Profile Description` and `Location` are

[1] The SemEval Top Participant result in Table 2 was restricted to our test set users to allow direct comparisons.

	SemEval F1
Most Frequent Class	67.0
Tweet Features Only	
SemEval Top Participant	68.4
User Attributes Only	
Name	66.5
Profile Description	64.5
Profile Picture	55.7
Location	67.8
Followees	72.3†‡
Inferred Factors	68.6
All User Attributes	69.5

Table 2: Comparison of different models for predicting the stance of a tweet. Models trained only on user attribute features perform as well as — or better than — models trained on features extracted from the tweet itself. SemEval F1 is the unweighted average between $F_{against}$ and F_{favor}. This version, which is the official metric used for evaluating SemEval participants, does not directly include the performance of the models on the *neutral* stance class. Statistical significance ($p < 0.05$) is indicated in comparison with the MFC (†) and the SemEval Top Participant (‡).

useful in four of the five tasks, excepting Climate. Followees and Inferred Factors are useful in all tasks, with Followees being more useful in three tasks and Inferred Factors more useful in two tasks.

User attributes, with the exception of Name and Profile Picture, predict stance better than MFC on average. For every target there is some user attribute that predicts stance better than MFC.

User attributes improve all targets: Profile Description, Location, and Followee information all provide improvements over the MFC for all targets. Name, which encodes inferred information about the ethnicity, nationality, and gender of the users, shows improvements for Hillary. The Profile Picture features carry some information for Feminism, Hillary and Abortion targets. These show that publicly available information about the users carry useful signals about the users' stances.

Inferred factors versus other features: We see substantial gains with Inferred Factors for Atheism, Feminism, and Abortion (at least 5 points in F1) but only minor gains for the Climate and Hillary targets. No single factor provides consistent gains across all targets. For instance, Personality is useful for Atheism and Feminism but not for Abortion, whereas Political Ideology is useful for Atheism and Abortion but not for Feminism. These show the importance of considering multiple factors.

We can drill deeper into personality factors and consider the correlations of personality dimensions with stances. Figure 1 provides the correlations of each dimension with stance. As can be seen in the figure, Atheism has the strongest correlations which explains the big prediction gains we see on that target. Although Climate and Feminism have a similar range of correlations, we don't see gains for Climate while we do for Feminism. This is likely due to the extreme class imbalance for Climate; the *against* class only made up 6% of tweets in the test set, and indeed few participants of the SemEval competition were able to beat the MFC for this target (Mohammad et al., 2016).

Overall, inferred factors perform better for Feminism and Abortion targets, while direct user attributes perform better for Atheism, Climate, and Hillary. These results suggest that stances on some targets correlate with psychological attributes such as personality and political orientation, whereas others are more correlated with demographic factors such as location.

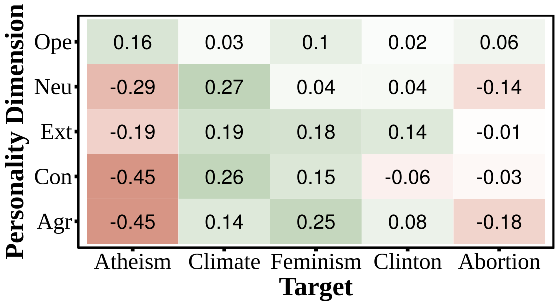

Figure 1: Pearson R correlation matrix for personality and stance.

Figure 2 shows the probability distribution of stances for Atheism over demographic variables. Age by itself has different distributions in *favor* and *against* populations, whereas there is no difference in the gender score distributions. Together, age and gender show stark contrasts in the distributions. Also note that for both gender and age, the *neutral* class distributions seem to capture a fairly symmetric split of *favor* and *against*. This may be related to the idea that stance is more of a user-level attribute than a message-level one, in that the *neutral* population actually contains users whose "real" stances are *favor* or *against* but

	F1					
	Atheism	*Climate*	*Feminism*	*Hillary*	*Abortion*	*All (Avg.)*
Most Frequent Class	61.7	60.8	45.8	47.0	55.8	54.2
User Attributes Only						
Name	61.7	59.1	46.1	48.8	55.5	54.2
Profile Description	64.3	58.6	**53.8**†	56.0†	59.6	58.5†
Profile Picture	58.5	57.6	47.0	50.1	57.3	54.1
Location	64.9	51.8	51.0	53.1†	**61.1**†	56.4
Followees	**73.2**†	**67.1**	52.4	**58.3**†	58.0	**61.8**†
Inferred Factors Only						
Demographics	61.9	60.5	49.6	46.8	55.8	54.9
Personality	**69.3**†	59.8	53.1†	47.0	55.8	57.0†
Political Ideology	65.8†	60.8	44.1	47.0	60.7†	55.7†
User Embeddings	64.5	59.0	43.5	47.9	56.0	54.2
All Inferred	67.5	**61.5**	**55.2**†	**48.9**	**63.4**†	**59.3**†

Table 3: Performance of stance prediction models trained only on user attributes, shown here for each of the different stance targets. Bold indicates best in column for user attributes and inferred factors. The weighted F1 is shown for each target and the last column is the unweighted average across all targets. † indicates statistical significance at the 0.05 level compared to the MFC baseline.

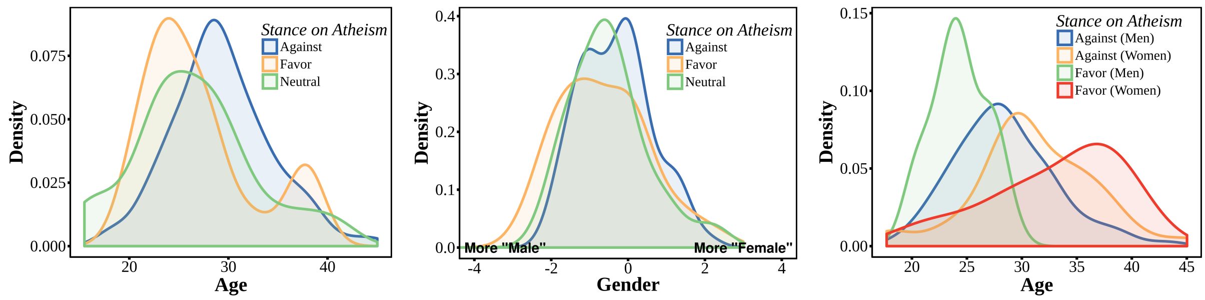

Figure 2: The first two graphs show the probability density of age (left) and gender (middle) for each of the three stance classes on the Atheism target. The rightmost graph shows the probability density of users' ages for Atheism, broken down by gender and class label (excluding *neutral*). There is a clear relationship between age, gender, and stance, demonstrating the need for user attributes.

which aren't expressed in the tweet itself. Overall, the plots show the degree to which stances can be separated simply by demographics but also suggest that one might benefit from variables capturing a combination of age and gender.

5 When is the Tweet Useful?

Tweets provide the most direct expression of a user's intent. However, the amount of task-related information in a tweet and the ability of tweet-derived features to model it reliably vary with the task. Table 4 compares tweet features and user attributes for stance, sarcasm, sentiment, and PP-attachment.

Overall, combining user attributes with tweet-derived features provides the best results for stance, sarcasm and sentiment.

Stance: Even though user attributes outperform tweet features when only considering *favor* and *against* classes, we find that tweet features turn out to be better when considering all three classes including the *neutral* class. The average F1 across all three classes for the tweet-only baseline is higher than any of the user attribute-only models (+1 point in average F1 over `Followees`, the best user attribute feature).

A closer look reveals why this is the case. For any given user, their positive or negative stance towards a target seldom changes. What may change instead is whether they express their stance or remain neutral when writing a particular tweet.

		F1			Acc.
		Stance	*Sarcasm*	*Sentiment*	*PP-Attach.*
Baselines	MFC	54.2	51.2	28.0	64.4*
	Tweet Only	62.8	74.1	69.2	**71.0**
No Tweet	Name	54.2	59.0†	38.9†	64.4
	Profile Description	58.5†	64.7†	40.3†	64.4
	Profile Picture	54.1	61.4†	40.7†	64.4
	Location	56.4	58.3†	38.7†	64.4
	Followees	61.8†	73.1†	42.3†	64.4
	Inferred Factors	59.3†	71.9†	41.4†	64.4
	All User Attributes	60.8†	74.4†	42.5†	64.4
With Tweet	Name + Tweet	62.9†	74.9†‡	69.4†	71.0
	Profile Description + Tweet	63.5†	73.4†	68.9†	71.0
	Profile Picture + Tweet	62.2†	71.1†	68.7†	71.0
	Location + Tweet	64.1†	74.0†	68.9†	71.0
	Followees + Tweet	**65.9**†‡	**78.6**†‡	**69.5**†	71.0
	Inferred Factors + Tweet	65.1†‡	77.3†‡	69.3†	71.0
	All User Attributes + Tweet	63.8†	76.8†‡	67.4†	71.0

Table 4: Using user attributes to predict stance, sarcasm, sentiment, and PP-attachment. Bold indicates best in column. Statistical significance ($p < 0.05$) is indicated in comparison with the MFC (†) and the tweet-only model (‡). *MFC computed by training a model only on the distance between the preposition and the candidate head.

Thus, while user attributes are better at predicting a user's overall stance, the tweet features provide a better indication of whether there is an expression of it in the specific tweet. Indeed, combining tweet features and user attributes yields additional gains in most cases: `Profile Description` (+0.7 points), `Location` (+1.3), `Inferred Factors` (+2.3), and `Followees` (+3.1). When combining `Followees` with tweet features, we see an 18.4 point improvement for $F_{neutral}$ on average over using `Followees` alone.

There can be non-linear interactions between the user attributes and the tweet features. For instance, we find that with a random forest classifier we can obtain a baseline performance of 65.0 F1 for the tweet-only features, which increases to 66.4 when combined with all user attributes. This exploration is beyond the scope of our work; here we only intend to show that even a simple combination can provide gains.

Sarcasm: Tweet features are no better than the combined set of user attributes for sarcasm, showing once again the extent of predictive power in user information.

`Inferred Factors` and `Followees` are the strongest user attributes and boost performance when combined with the tweet features, providing roughly 3.2 and 4.5 point gains respectively. `Name` embeddings which carry nationality, ethnicity, and gender information provide a 0.8 point gain. The other features provide no gains when combined with tweet features. Combining all user attributes performs worse than using `Followees` or `Inferred Factors` alone, presumably due to pushing the bounds in terms of total number of features given limited observations.

Sentiment: User attributes appear far less useful than tweet features for sentiment. While users can lean positive or negative overall, sentiments are contextual and are best inferred from expressions in the tweet. Still, combining user attributes with tweet features yields minor gains.

PP-Attachment: The user attributes provide no useful predictive value. They do not do better than even the simple MFC baseline[2] and combining with text doesn't provide any improvements, reflecting the idea that this task is closer to something purely linguistic. Even so, prior work suggests user attributes can still benefit PP-attachment when using more sophisticated approaches like user-factor adaptation (Lynn et al., 2017).

[2] For PP-attachment, the MFC is computed by training the model only on the distance between the preposition and the candidate head.

5.1 Trait versus State

Overall, we see that stance and sarcasm benefit most from user attributes, sentiment benefits a little, and PP-attachment not at all. Supported by these results, we theorize that outcomes which are more "trait-like" benefit more from user attributes than those that are more "state-like". Trait-like outcomes are those that tend to be stable over time, such as stance; while the exact expression may vary from tweet to tweet, a person's overall stance is likely to remain relatively unchanged across many messages. State-like outcomes, on the other hand, are those that change frequently, such as sentiment. Sarcasm is somewhere in between — trait-like, in that a person can have a predisposition for being sarcastic, but the expression at message level still largely depends on context. PP-attachment is a state-like outcome as it depends entirely on the syntactic structure of the tweet.

6 Discussion

We found: (1) state-of-the-art tweet stance detection can be achieved without even using the tweet and instead using user attributes; (2) user attributes have varying predictive utility depending on the target of stance (e.g. atheism versus abortion); (3) different types of user attributes are valuable for different tasks — out of those we considered, followees on Twitter were most valuable; and (4) adding the tweet content back in on top of user attributes yields even greater performance.

The fact that user attributes predict stance better than tweet attributes may be surprising considering that the gold-standard labels were done by human annotators who were not privy to user attributes of the tweet author (Mohammad et al., 2016). Annotators were in fact trying to guess what the user's stance was from their tweet. They were instructed to "infer from the tweet that the tweeter [supports|is against|has a neutral stance] towards the target" (or that it was not possible to tell). However, our predictive models without the tweet were not even seeing the same information as these humans they were trying to mimic, and yet these tweetless models predicted just as well as models that did see the tweet. This raises interesting questions about whether the tweet-based models are unable to reliably use the information in text or whether the annotators used implicit signals in tweets to infer user attitudes towards the target. Joseph et al. (2017) raise a similar issue with systematic errors in stance annotation according to the context provided to human annotators.

This raises a counterpoint to the standard framing of social media tasks as making inferences over text alone. These results, combined with the fact that similar patterns were replicated with sarcasm and sentiment, speak to the question: **How much merit is there in attempting social media tasks agnostic to user attributes?**

For applications of stance, sarcasm, and sentiment tasks, such as tracking changes over time (stance, sentiment), or identifying particular tweets to interpret differently (sarcasm), it would certainly be less than ideal to simply predict the same outcome for every tweet from a given user as our *tweetless* models would do. Thus, we can at least say that there is value in the tweet or individual document itself, so the question is how to integrate user attributes and the tweet. Prior work on user-factor adaptation (Lynn et al., 2017) and use of residualized models (Zamani and Schwartz, 2017; Zamani et al., 2018) provide interesting avenues for exploration.

The results provide some insights into designing future social media tasks. First, given the strong impact of user attributes on these tasks, it becomes readily apparent that the diversity of the user base is a key consideration in designing these tasks. Consider a training sample of tweets that is drawn only from users with certain attributes. Not only will the test performance on other users suffer, we also lose the opportunity to leverage strong user-level correlations in making predictions. A secondary implication is that when considering performance of user attributes on these tasks, care must be taken to see whether there is a representative diversity in the training sample before dismissing the value of the attribute.

We also propose that shared tasks consider user attribute baselines, mirroring the idea of "controls" in social scientific studies, whereby the goal is to predict above and beyond such attributes or leverage both most effectively. Setups like this have been done for some user-level tasks, such as providing age and gender estimates for mental health prediction (Coppersmith et al., 2015) or socioeconomic information for assessing community life satisfaction (Schwartz et al., 2013) or market prices (Zamani and Schwartz, 2017). However, for document-level tasks like those Twitter tasks we explore, a comparison to user attributes has usually been restricted to case studies such as those we mentioned in Section 2.

Still, it can be challenging to determine what user attributes to include as a baseline. Other fields, such as psychology, suggest always controlling for basic human traits — such as age and gender — as well as theoretically-related variables such as socioeconomic variables or, perhaps, political ideology in the case of stance detection (Gazzaniga and Heatherton, 2015). Another approach could be to consider what other information is readily available — those we have included here are typically available if one's documents are tweets but, for example, one also might often find location, demographics, and years of experience available for news or scientific articles.

7 Conclusion

More and more natural language processing tasks focus on social media. With advances in incorporating user information it has become increasingly clear that many tasks are best framed in user and social contexts. This work emphasizes the increasingly prominent role for user attributes in language tasks. We have shown state-of-the-art performance in tweet stance detection without the tweet itself, and shown that stance classification, sarcasm detection, and sentiment analysis models can be significantly improved with user factors. We find variance in utility of different user attribute features across tasks and raise important practical considerations for designing future social media tasks and their solutions.

References

Silvio Amir, Byron C Wallace, Hao Lyu, Paula Carvalho, and Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *CoNLL 2016*, page 167.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on Twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Yonatan Belinkov, Tao Lei, Regina Barzilay, and Amir Globerson. 2014. Exploring compositional architectures and word vector representations for prepositional phrase attachment. *TACL*, 2:561–572.

Wei-Fan Chen and Lun-Wei Ku. 2016. Utcnn: a deep learning model of stance classificationon on social media text. In *Proceedings of COLING: Technical Papers*, pages 1635–1645.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *CLPsych @ North American Association for Computational Linguistics*, pages 31–39.

Michael Gazzaniga and Todd Heatherton. 2015. *Psychological Science: Fifth International Student Edition*. WW Norton & Company.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *IJCNLP*, pages 1348–1356.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of ACL*.

Dirk Hovy and Tommaso Fornaciari. 2018. Increasing in-class similarity by retrofitting embeddings with demographic information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 671–677.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of KDD*, pages 217–226. ACM.

Kristen Johnson and Dan Goldwasser. 2016. "All I know about politics is what I read in Twitter": Weakly supervised models for extracting politicians' stances from Twitter. In *Proceedings of COLING*, pages 2966–2977.

Kenneth Joseph, Lisa Friedland, William Hobbs, David Lazer, and Oren Tsur. 2017. Constance: Modeling annotation contexts to improve stance classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1126–1135.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Writer profiling without the writers text. In *International Conference on Social Informatics*, pages 537–558. Springer.

Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. 2015. Your sentiment precedes you: Using an authors historical tweets to predict sarcasm. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 25–30.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Rebecca Knowles, Josh Carroll, and Mark Dredze. 2016. Demographer: Extremely simple name demographics. In *EMNLP Workshop on NLP and Computational Social Science*, pages 108–113. Association for Computational Linguistics.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of EMNLP*.

Vivek Kulkarni, Margaret L Kern, David Stillwell, Michal Kosinski, Sandra Matz, Lyle Ungar, Steven Skiena, and H Andrew Schwartz. 2017. Latent human traits in the language of social media: An open-vocabulary approach. *arXiv preprint arXiv:1705.08038*.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? Quantifying the geographic variation of language in online social media. In *Tenth International AAAI Conference on Web and Social Media*.

Veronica E. Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Empirical Methods in Natural Language Processing*, pages 1146–1155.

Travis Martin, Jake M Hofman, Amit Sharma, Ashton Anderson, and Duncan J Watts. 2016. Exploring limits to prediction in complex social systems. In *Proceedings of the 25th International Conference on World Wide Web*, pages 683–694. International World Wide Web Conferences Steering Committee.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of SemEval-2016*, volume 16.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of SemEval-2013*.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter.

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 729–740.

Maarten Sap, Gregory J. Park, Johannes C. Eichstaedt, Margaret L. Kern, David Stillwell, Michal Kosinski, Lyle H. Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of EMNLP*.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, and Lyle Ungar. 2013. Characterizing geographic variation in well-being using tweets. In *International Conference on Web blogs and Social Media*.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 116–125.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327–335.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP*.

Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012a. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596. Association for Computational Linguistics.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012b. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Yi Yang and Jacob Eisenstein. 2015. Putting things in context: Community-specific embedding projections for sentiment analysis. *CoRR*, abs/1511.06052.

Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, and Steven Skiena. 2017. Nationality classification using name embeddings. In *CIKM*, pages 1897–1906.

Mohammadzaman Zamani, H Andrew Schwartz, Veronica E Lynn, Salvatore Giorgi, and Niranjan Balasubramanian. 2018. Residualized factor adaptation for community social media prediction tasks. *arXiv preprint arXiv:1808.09479*.

Mohammadzaman Zamani and Hansen Andrew Schwartz. 2017. Using twitter language to predict the real estate market. In *EACL 2017: European Association for Computational Linguistics*, page 28.

Guido Zarrella and Amy Marsh. 2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection. *Proceedings of SemEval*, pages 458–463.