Identifying Effective Signals to Predict Deleted and Suspended Accounts on Twitter across Languages

Svitlana Volkova and Eric Bell
Data Sciences and Analytics, National Security Directorate
Pacific Northwest National Laboratory
902 Battelle Blvd, Richland, WA USA 99354

Abstract

Social networks have an ephemerality to them where accounts and messages are constantly being edited, deleted, or marked as private. This continuous change comes from concerns around privacy, a potential desire for to be forgotten and suspicious behavior. In this study we present a novel task – predicting suspicious e.g., to be deleted or suspended accounts in social media. We analyze multiple datasets of thousands of active, deleted and suspended Twitter accounts to produce a series of predictive representations for the removal or shutdown of an account. We selected these accounts from speakers of three languages – Russian, Spanish, and English to evaluate if speakers of various languages behave differently with regards to deleting accounts. We compared the predictive power of the state-of-the-art machine learning models to recurrent neutral networks trained on previously unexplored features. Furthermore, this work is the first to rely on image and affect signals in addition to language and network to predict deleted and suspended accounts in social media.

We found that unlike widely used profile and network features, the discourse of deleted or suspended versus active accounts forms the basis for highly accurate account deletion and suspension prediction. More precisely, we observed that the presence of certain terms in tweets leads to a higher likelihood for that user's account deletion or suspension. Moreover, despite image and affect signals yield lower predictive performance compared to language, they reveal interesting behavioral differences across speakers of different languages. Our extensive analysis and novel findings on language use and suspicious behavior of speakers of different languages can improve the existing approaches to credibility analysis, disinformation and deception detection in social media.

Introduction

Large volumes of personalized, timely, diverse, and multilingual data from social media services like Facebook, Twitter, Google+, Youtube etc., have been successfully used to answer various social science, political science, computational linguistics, and sociolinguistics questions. These questions include, but are not limited to real-world event detection (Atefeh and Khreich 2015), user-centric analytics (Preoţiuc-Pietro et al. 2015), personality detection (Golbeck et al. 2011), discourse analysis (Eisenstein et al. 2014; Kim et al. 2014), emotion detection and analysis (Volkova and Bachrach 2015), large scale passive polling (O'Connor et al. 2010), monitoring sports events (Kim et al. 2015) and health analytics (Paul and Dredze 2011).

These studies draw conclusions from large samples of social media data. However, communications sampled from public social media might be deleted over time by users, or removed by social media service providers after being flagged as those being produced by social bots, spam accounts, fraudulent accounts that can potentially be spreading misinformation, deceptive content and propaganda. According to the Twitter spam policy,[1] users may not use Twitter for the purposes of spamming. Accounts that violate this rule may be suspended or terminated, and their tweets removed from Twitter. Other reasons for account suspension include:

- account security at risk e.g., if an account has been hacked or compromised;
- abusive tweets or behavior e.g., sending threats to others or impersonating other accounts.

Moreover, recently Facebook revealed that 83 million of its users may be fake.[2] Another report claims that 1 in 10 accounts on Twitter is fake.[3] Computational social science experiments conducted on samples that contain suspicious accounts are based on data that is modified over time by user deletions or suspensions, resulting in future researchers being unable to construct the same data and reproduce the same results. Further verification on the representativeness, bias, and validity of these studies is a recognized need (Wallach 2014; Tufekci 2014; Bamman 2016).

Furthermore, early detection of deleted and suspended accounts that can potentially be spreading misinformation and deceptive content is extremely important to ensure a safer and healthier environment in social media.[4]

In this work we propose a technique to automatically predict "to be deleted accounts" (both suspended and intentionally deleted by users) on Twitter with the goal of excluding these accounts from sampled data to improve reproducibility of future studies.

[1] Twitter spam policy: https://support.twitter.com/articles/18311

[2] http://www.bbc.com/news/technology-19093078

[3] http://www.nbcnews.com/business/1-10-twitter-accounts-fake-say-researchers-2D11655362

[4] http://www.theatlantic.com/international/archive/2014/08/the-kremlins-troll-army/375932/

Our main contributions include:

- Building predictive models and **contrasting the effectiveness of different representations** – e.g., language, network structure, images, opinions and emotions for detecting deleted and suspended accounts on Twitter.
- Comparing the state-of-the-art models with recurrent neural networks for predicting suspicious behavior across **multiple languages – Russian, Spanish, and English**.
- Assessing the predictive power of previously **underexplored signals in social media – images and affects** expressed in tweets for detecting suspicious accounts.

To the best of our knowledge, this is the first work that compares the predictive power of recurrent neural networks with the state-of-the-art machine learning models e.g., logistic regression for account deletion prediction. Moreover, unlike the existing work on social bot prediction (Ferrara et al. 2016) and suspended account analysis (Thomas et al. 2011), we perform deep linguistic analysis of user-generated content to contrast the predictive power of features across three languages, including those that have never been used for account deletion prediction such as: opinions, emotions, word embeddings, topics, and images, in addition to well-studied profile, network, and behavior signals.

The proposed models rely on a *limited amount of user content*, and, thus, are capable of making predictions in a constrained-resource scenario e.g., with only 20 tweets per user. By relying on topic and embedding features, our models make predictions from a *low-dimensional feature space*, and, therefore, are capable of processing high volumes of streaming data very fast with low memory requirements. In addition, we experiment with recently emerged *neural network approaches* and compare their predictive power to the state-of-the-art classifiers. Finally, our models *do not rely on language-specific resources* and perform well across languages, including morphologically rich languages like Russian and Spanish.

Related Work

The existing body of work on detecting suspicious accounts in social media focuses on social bot prediction (Hwang, Pearce, and Nanis 2012; Chu et al. 2012; Ferrara 2015), sybil detection (Cao et al. 2012), suspended account analysis (Thomas et al. 2011), and spam detection (Lin and Huang 2013; Guo and Chen 2014; Yang et al. 2012; Zhang et al. 2014). In a recent survey on social bot detection Ferrara et al., 2016 divides proposed approaches into three types that: (1) rely on a social network structure[5] (Cao et al. 2012), (2) take advantage of crowdsourcing (Wang et al. 2012), and (3) use machine learning to estimate predictive features as discussed below.

Influence Bot Detection The Defense Advanced Research Projects Agency's (DARPA) Twitter bot challenge focused on detecting influence bots (Subrahmanian et al. 2016). Interestingly, their best performing system relied on shallow linguistic and sentiment features (Dickerson, Kagan, and Subrahmanian 2014). Recent work on influence bot detection in RuNet (Lawrence 2015) analyzed 20,500 Twitter accounts that tweeted similar statements around key breaking news and events.

Spam Account Detection Guo and Chen, 2014 proposed a supervised model that relies on shallow content-based features (aka stylistic features) e.g., proportion of hashtags, mentions, retweets and geographic features to classify spam accounts on manually-labeled datasets. Similarly, Lin and Huang, 2013 relied on two behavior features – URL rate and interaction rate to detect spam accounts. Lee, Eoff, and Caverlee, 2011 used tweeting activity, following strategies, behavior over time, and links to classify content polluters on Twitter e.g., duplicate spammers and malicious promoters.

Suspended Account Analysis The most similar work to ours is by Thomas et al. 2011, where the authors analyzed 1.1 million accounts suspended by Twitter to characterize spammers' behavior and their lifetime activity. However, they have not built predictive models for distinguishing deleted accounts from suspended and active accounts, or performed linguistic analysis of user content.

Tweet Deletion Prediction Interestingly, linguistic features have only been used to predict malicious or deleted tweets rather than deleted accounts. For example, researchers built language models classify deleted vs. non-deleted tweets (Martinez-Romo and Araujo 2013). Almuhimedi et al., 2013 focused on keyword analysis, temporal aspects of deleted tweets, and their geo-tagged information. Some approaches relied on social, author, and text features (Petrovic, Osborne, and Lavrenko 2013). Potash, Bell, and Harrison, 2016 used topic features and character-level embeddings to predict deleted tweets. Others analyzed millions of messages from Weibo and Twitter for political censorship and found that the presence of politically sensitive terms in messages leads to anomalously higher rates of deletion (Bamman, O'Connor, and Smith 2012).

To the best of our knowledge, there is no existing work that contrasts the predictive power of the state-of-the-art classifiers learned from both shallow and deep linguistics features with neural network models for account deletion prediction on Twitter across multiple languages.

Data

English and Spanish Data Collection

Data for English and Spanish tweet deletion seed materials was selected from an archive of the public 1% Twitter feed with no filtering criteria. The time period covered was September 1, 2015 through December 30, 2015.

After issuing a query for tweets in the target language in January 2016, batches of 100 unique users were queried against the public Twitter API. Those returning active profiles were classified as *non-deleted users*. Missing profiles were classified as *deleted users*. Once approximately $DS = 100{,}000$ unique non-active users were encountered per language, further queries were issued against the original dataset to retrieve all tweets in the repository by those

[5] SybilRank algorithm identifies densely interconnected groups of sybils (groups of social bots).

users. Moreover, we queried Twitter API to further verify whether the account was deleted by a user or suspended. Selecting randomly from within the sample of non-deleted users, and retaining only individuals with at least 5 tweets in our dataset, we extracted another $\bar{D} = 100{,}000$ unique non-deleted users. Examples of the types of content in deleted user tweets include – *"... best herbs for weight loss begin with green tea..."* and *"...lo mucho que quiero estar en tu corazon tatuado ... " (how much I want to be in your heart tattooed ...)* Examples shown have been selected to show generalities, but are not actual deleted tweets in adherence to Twitter policy and user privacy.

Russian Data Collection

We sampled Twitter accounts which mention Russia-Ukraine crisis-related keywords in Russian or Ukrainian (Volkova et al. 2016). The example tweet content (translated) with crisis-relevant discourse – *Cyborgs hung the Ukrainian flag in Donetsk Airport.*

The original dataset had 3.5 million users who used crisis-relevant keywords during this period. We then re-crawled a random sample of 1 million accounts within a couple of months (Jun 2015) of the initial data collection (Mar 2015). We discovered that 30% of previously active accounts were not active anymore (have been deleted or suspended). We re-crawled these accounts in Dec 2015 to validate the accounts that have been deleted or suspended as of Mar 2015 and still remain non-active as of Dec 2015. We call this portion of the data *deleted and suspended accounts* $DS = 94,170$. We then randomly sampled the same number of accounts that were still active e.g., not deleted as of Mar 2015 and still remain active as of Dec 2015. We call this portion of the data *non-deleted accounts* $\bar{D} = 94,170$. For each user $u \in \{D, S, \bar{D}\}$ or $u \in \{DS, \bar{D}\}$ we were able to access at least 20 tweets as well as user profile metadata.

In Table 1 we present statistics for English, Spanish, and Russian datasets in terms of the total number of tweets per language within deleted (D), suspended (S) and non-deleted ($\bar{D}$) accounts, and the average numbers of tweets per user.

Type	Mean	Tweets	Accounts
ENGLISH			
Deleted D	18	1,479,747	82,435
Suspended S	68	1,200,257	17,565
Non-deleted $\bar{D}$	35	3,503,232	100,000
SPANISH			
Deleted D	9	855,751	91,161
Suspended S	14	121,935	8,839
Non-deleted $\bar{D}$	130	12,999,202	100,000
RUSSIAN			
Deleted D	20	275,275	13,845
Suspended S	20	1,601,483	80,325
Non-deleted $\bar{D}$	20	1,872,723	94,170

Table 1: The number of deleted D, suspended S and non-deleted $\bar{D}$ accounts and tweets per language.

Approach

We experiment with thee types of models for account deletion prediction – deleted vs. suspended (2-way: D-S), deleted+suspended vs. non-deleted (2-way: DS-ND), and deleted vs. suspended vs. non-deleted (3-way: D-S-ND).

Models We used scikit-learn toolkit (Pedregosa et al. 2011) to build models that can distinguish between deleted, suspended and non-deleted accounts. We tested several models including SVM and Random Forest. However, they yield lower performance compared to log-linear models and excluded them from our analysis.

In Table 2 we outline a comprehensive list of features we used to build models for account deletion prediction by significantly expanding the list of features that have been previously used for bot detection on Twitter. In addition to previously used account and behavior features, our models rely on network structure features and deeper linguistic analysis of tweets generated by users, including topics, embeddings, as well as novel image and affect features extracted from user communications.

Recurrent neural networks have been extensively used for sentence classification recently.[6] Following standard practices for sentence classification, we implement a Long Short-Term Memory neural network (Hochreiter and Schmidhuber 1997) in Keras[7] for binary and multi-class classification. We use an embedding layer, a recurrent layer and an output layer. We rely on sigmoid activation function[8] and learn weights using RMSprop optimizer.[9] We contrast LSTM performance with the state-of-the-art log-linear models learned from features discussed below.

Tweet Ngrams Russian and Spanish are morphologically rich languages. To reduce sparsity and ensure better model generalization, we lemmatized words using the pymorphy2 package[10] for Russian and snowball stemmer[11] for Spanish. We started by extracting ngram features from the preprocessed lemmatized tweets. We then excluded all stopwords and words with frequency less than five. We ran our experiments with log-linear models by varying word ngram size (unigrams, bigrams, and trigrams) for binary vs. normalized frequency-based ngram features. Below we describe how we extract embedding, topic, affect and novel image features from user communications across languages.

Tweet Ngrams + LSA We performed linear dimensionality reduction on feature vectors extracted using normalized frequency-based bigram features as described above using Latent Semantic Analysis (LSA) (Dumais 2004) implemented as truncated Singular Value Decomposition (SVD) in scikit-learn.[12] Similarly, we performed linear dimen-

[6] karpathy.github.io/2015/05/21/rnn-effectiveness/
[7] https://keras.io/getting-started/sequential-model-guide/
[8] https://keras.io/activations/
[9] https://keras.io/optimizers/#rmsprop
[10] https://pypi.python.org/pypi/pymorphy2
[11] https://pypi.python.org/pypi/snowballstemmer
[12] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

PROFILE FEATURES $\lvert f^{prof} \rvert = 12$
days since account creation, number of followers, number of friends, number of favorites, number of tweets, friend-to-follow ratio, name length in chars, bio in chars, screen name length in chars, screen name length in words, bio length words, avg. number of tweets per hour
SYNTACTIC AND STYLICTIC FEATURES $\lvert f^{syn} \rvert = 14$
aver. tweet length in words, aver. tweet length in chars, retweet rate: prop. of RTs to tweets, uppercase word rate, elongated word rate, repeated mixed punctuation rate, prop. of tweets with links, tweets that are retweets (RTs), prop. of tweets with mentions, hashtags, punctuation, emoticons, mention, hashtag, url rate per word
LEXICAL FEATURES
Tweet ngrams (binary vs. count-based)
Tweet ngrams + LSA with $c = [50, \ldots, 1000]$
Topics with $t = [50, \ldots, 1000]$ topics
Embeddings with $d = [30, 50, 100 \ldots 2000]$
NETWORK FEATURES
Mentions (binary vs. count-based unigrams)
Mentions + LSA with $c = [50, \ldots, 1000]$
Hashtags (binary vs. count-based unigrams)
Hashtags + LSA with $c = [50, \ldots, 1000]$
AFFECT FEATURES $\lvert f^{affect} \rvert = 10$
Proportion of tweets with six Ekman's emotions (joy, sad, fear, disgust, anger, surprise), Proportion of tweets with positive, negative and neutral sentiments
IMAGE FEATURES $\lvert f^{image} \rvert = 2048$
Image representation 2048-dim vector extracted using CNN

Table 2: Profile, syntactic, stylistic, lexical, network, and affect features for account deletion prediction.

sionality reduction on feature vectors with hashtags and @mentions. We varied the number of dimensions $c = [50, 100, 500]$ to get the best prediction performance and report the results for $c = 100$ dimensions.

Topics We learned topics using LDA[13] on an independent sample of one million tweets for each language (Blei, Ng, and Jordan 2003). We varied the number of topics $t = [50, 100, 250, 500, 1000]$, and tuned Dirichlet priors α and β. We found that high values of alpha made each tweet contain all topics (or the majority of the topics) rather than a single topic. Thus, topics are less interpretable. Low values of alpha made each tweet contain a mixture of a few topics or even a single topic. Thus, topics are more recognizable. Similarly, high values of beta made each topic contain a mixture of all of the words, and low values of beta made each topic include a mixture of just a few words (less variance within the topic). We found that the optimal values of priors are $\alpha = 0.1$ and $\beta = 0.005$, and topics $t = 1000$ by maximizing log-likelihood on a development subset of tweets.

Embeddings Text embeddings represent words as numeric vectors in a continuous space, where words within similar contexts appear close to one another. The majority of NLP applications are using word embeddings as features for downstream prediction tasks e.g., part-of-speech tagging (Santos and Zadrozny 2014), named entity recognition (Passos, Kumar, and McCallum 2014) and dependency parsing (Lei et al. 2014).

For English we relied on pre-trained embeddings obtained using GLoVe[14] (Pennington, Socher, and Manning 2014), Normalized Pointwise Mutual Information (NPMI) (Lampos et al. 2014) and Word2Vec[15] (Mikolov et al. 2013). For Russian and Spanish we learned word embeddings using Word2Vec model implemented in the gensim package with a layer size of 50. The embeddings are learned on the same corpus of one million tweets as LDA topics. After learning embeddings, we assigned words to clusters by measuring cosine similarity between embedding pairs, and computed clusters using spectral clustering over a word-to-word similarity matrix.

Opinions and Emotions To extract sentiment features for Russian we predict a polarity score for every tweet per user using the state-of-the-art sentiment classification system for Russian (Chetviorkin, Moscow, and Loukachevitch 2014). Polarity scores vary around 0 (neutral) between -2 (negative) and +2 (positive). We then calculate mean polarity, and the proportions of positive, negative, and neutral tweets per account. To extract sentiment features for English and Spanish we predict sentiment labels – positive, negative, or neutral, for every tweet per user using pre-trained models from (Volkova and Bachrach 2015), respectively. We then calculate proportions of positive, negative, and neutral tweets per user account. To extract emotion features across all languages, we predict one of six Ekman's emotions – sadness, joy, fear, disgust, surprise, and anger for each tweet using an approach developed by (Volkova and Bachrach 2015). Similar to sentiment features, we use six emotion proportions as features.

We acknowledge that relying on pre-trained sentiment and emotion models is not optimal. However, these models for affect classification yield the state-of-the-art performance across languages (Volkova and Bachrach 2015; Chetviorkin, Moscow, and Loukachevitch 2014; Volkova, Wilson, and Yarowsky 2013).

Images Beyond just being a classification system, Convolutional Neural Networks (CNNs) can be used as feature extractors, whereas the features produced by the top layers of the CNN can be used with great efficacy on tasks not related to the original task that the network was trained on, referred to as transfer learning (Yosinski et al. 2014). In this work, we took a similar approach; we used the Inception v3 model trained on the ImageNet data set (Russakovsky et al. 2015). The top softmax layer was removed from the network, leaving the final fully connected layer, which produced a 2048-dimensional vector for each image in our data set.

[13]https://pypi.python.org/pypi/lda

[14]http://nlp.stanford.edu/projects/glove/

[15]https://radimrehurek.com/gensim/models/word2vec.html

Language	ENGLISH			RUSSIAN			SPANISH		
Feature Type — Task	D-S-ND	DS-ND	D-S	D-S-ND	DS-ND	D-S	D-S-ND	DS-ND	D-S
LOG-LINEAR									
Account + Behavior	0.65	0.75	0.82	0.78	0.85	0.86	0.65	0.72	0.90
Style + Syntax	**0.87**	0.62	0.87	0.72	0.81	0.86	0.60	0.64	0.90
Tweets	0.84	**0.88**	0.89	**0.82**	**0.87**	0.83	**0.74**	0.79	**0.94**
Tweets + LSA	0.79	0.84	0.86	0.79	0.84	0.85	0.67	0.75	0.90
Topics	0.79	0.83	0.87	0.77	0.81	0.83	0.74	0.76	0.91
Embeddings	0.81	0.86	**0.91**	0.72	0.76	**0.94**	0.73	**0.82**	0.87
Hashtags	0.68	0.77	0.85	0.67	0.76	0.84	0.64	0.71	0.92
Mentions	**0.72**	**0.79**	**0.86**	**0.69**	**0.78**	**0.85**	**0.63**	**0.72**	**0.92**
Hashtags + LSA	0.40	0.70	0.83	0.63	0.73	0.84	0.58	0.69	0.92
Mentions + LSA	0.58	0.70	0.84	0.64	0.72	0.85	0.55	0.68	0.92
Sentiment + Emotion	0.76	0.53	0.76	0.62	0.72	0.83	0.53	0.30	0.88
Images (CNN)	0.52	0.54	0.85	–	–	–	0.52	0.54	0.89
LSTM									
Tweets + Network	0.84	0.85	**0.95**	**0.90**	**0.92**	**0.98**	**0.79**	0.80	**0.96**

Table 3: Classification results (F1) obtained using log-linear models and neutral networks (LSTM) for deleted vs. suspended vs. non-deleted (D-S-ND), deleted + suspended vs. non-deleted (DS-ND), and deleted vs. suspended (D-S) accounts based on individual feature types: profile, lexical (tweet ngrams, tweets + LSA with $c = 100$, topics with $t = 1000$ and embeddings with $d = 2000$), network (mentions and hashtags), image representations and affects (sentiments and emotions).

Experimental Results

This section describes classification results obtained using our models learned from profile, language, network, and affect features, and quantitatively analyzes and contrasts the predictive power of different feature types.

In Table 3 we report classification results for deleted vs. suspended (D-S), deleted+suspended vs. non-deleted (DS-ND), and deleted vs. suspended vs. non-deleted (D-S-ND) tasks obtained using 10-fold cross-validation (c.v.) with different feature combinations across three languages. Relying on earlier findings by Lee, Eoff, and Caverlee 2011 on model robustness toward different training mixtures of content polluters (deleted and suspended) vs. legitimate users, we balanced our deleted vs. non-deleted account datasets (DS-ND) to simplify the interpretation of classification results. For the experiments with imbalanced classes e.g., D-S-ND and D-S we report weighted F1 score.[16] We found that depending on language, different feature types lead to different performance. In terms of previously understudied content features syntactic and stylistic features and tweet ngrams yield the best performance for English and Russian, and embeddings features for Spanish. We outline our detailed findings below.

Profile features yield higher performance in terms of F1 score for Russian but lower for English and Spanish (except for D-S classification). *Syntax and style features* show higher F1 for Russian (0.81) than for English (0.62) and Spanish (0.64) for DS-ND, and the best F1 for English (0.87) for D-S-ND and Spanish (0.90) for D-S classification.

Tweet ngrams demonstrate higher performance for English but lower F1 for Russian and Spanish (except for D-S classification). Network features yield comparable results across all languages. As expected, dimensionality reduction e.g., *LSA on mentions or hashtags*, yield lower F1 across all languages and classification tasks compared to the unigram features on hashtags and mentions. Even though LSA models are efficient to use in a streaming setting they tend to overgeneralize and predict a higher rate of false positives. When we apply other dimensionality reduction techniques e.g., *topics or embeddings* to user tweets, we observe lower performance compared to ngrams (except for Spanish DS-ND and for English and Russian D-S). It might be due to temporal differences in the vocabulary obtained when learning topics and embeddings vs. the main data used for the experiments. Overall, *embeddings* demonstrate much higher performance for English than for Spanish and Russian.

Sentiment and emotion features yield much higher performance for Russian (0.72) than English (0.53) and Spanish (0.30) for DS-ND classification; however, they demonstrate the best F1 for English (D-S-ND) and Spanish (D-S). *Image* features yield the lowest performance for DS-ND and D-S-ND classification, and comparable F1 for D-S classification for English and Spanish.[17]

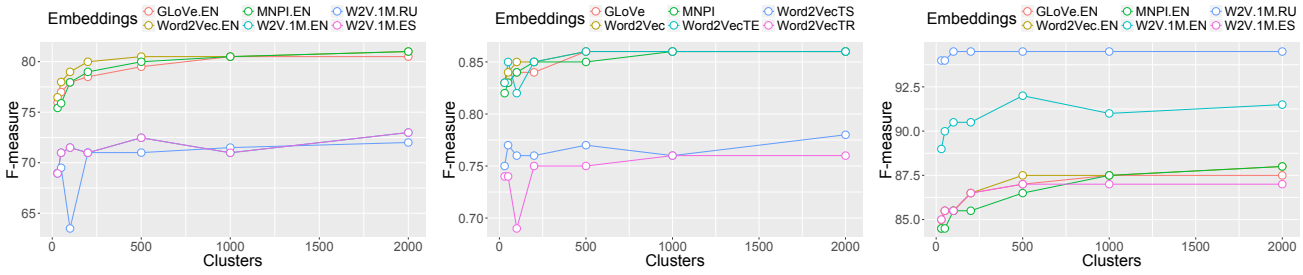
Table 3 also reports results obtained using LSTM models learned from tweet + network (hashtag and mention) features. We observe that neural network models consistently outperform log-linear models learned from different features for Russian. LSTMs yield the highest performance for deleted vs. suspended classification across languages, and comparable results for DS-ND and D-S-ND classification for English and Spanish. However, LSTMs take longer to train compared log-linear models – e.g., 30min per fold per classification task with 20 epochs on a single GPU.

[16]To find weighted F1 we calculate metrics for each label, and find their average, weighted by support (the # of true instances for each label). This alters macro F1 to account for label imbalance.

[17]We could not download images for the Russian data.

System	F-score	Model	Data
(Benevenuto et al. 2010)	0.87	SVM	55M (Test: 1K tweets)
(Martinez-Romo and Araujo 2013) – LM	0.88	SVM	20M (Test: 500K tweets)
(Petrovic, Osborne, and Lavrenko 2013) – BoW	0.27	SVM	68M (Test: 7.5M tweets)
(Potash, Bell, and Harrison 2016) – LDA, Embeddings	0.46	RF	90K (10-fold c.v.)
(Lee, Caverlee, and Webb 2010) – Syntactic, Behavior	0.89	RF	210K (Test: 1K users)
(Yang, Harkreader, and Gu 2011) – Syntactic, Tweet similarity	0.88	RF	500K (Test: 5.5K users)
This work (DS-ND) – Tweet Ngrams, Embeddings	0.92 (0.88)	LSTM (LL)	200K (Test: 35K; 10-fold c.v.)
This work (D-S) – Tweet Ngrams, Embeddings	0.98 (0.94)	LSTM (LL)	100K (Test: 17K; 10-fold c.v.)
This work (D-S-ND) – Syntax, Style, Ngrams	0.90 (0.87)	LSTM (LL)	200K (Test: 35K; 10-fold c.v.)

Table 4: Comparison of different account and tweet deletion prediction approaches on Twitter.



(a) 3-way: D-S-ND (b) 2 way: DS-ND (c) 2 way: D-S

Figure 1: Classification performance using different embedding types (GLoVe, Word2Vec, NMPI) with log-linear models.

Pre-trained Embedding Evaluation In Table 3 we report account deletion prediction results obtained using 2000 embedding clusters that lead to the best performance. In Figure 1 we show how varying the number of clusters from 30 to 2000 and embedding type e.g., GLoVe (Pennington, Socher, and Manning 2014), NPMI (Lampos et al. 2014), and Word2Vec (Mikolov et al. 2013) influences classification performance. We found that all types of embeddings learned for English yield higher F1 scores compared to embeddings learned from Spanish and Russian (except for D-S classification). Embeddings learned using Word2Vec outperforms NPMI and GLoVe when the number of word clusters is less than 1000. We also observe that increasing the number of clusters leads to better performance. We find similar trends when we vary the number of topics.

Comparison to Other Systems Table 4 shows the comparison of the most similar systems for account and tweet deletion prediction to ours – for example, the approaches that rely on shallow (syntactic) or deep linguistic features. However, the comparison is not fair because the data each system was evaluated on is different in each case and we evaluate our models on much larger datasets (200K users) across three languages. These results suggest that language features are indeed useful for account deletion prediction, and that classification performance varies depending on user language, as well as the type and the size of a data sample.

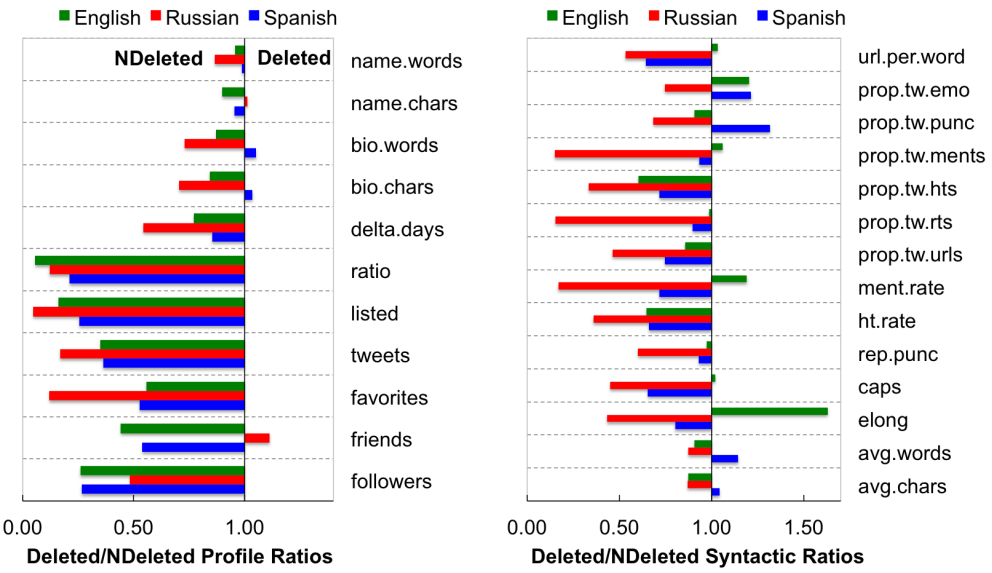
Contrastive Analysis

To show that differences between deleted + suspended (DS) and non-deleted (ND) accounts are statistically significant we performed Mann-Whitney tests on account, affect, and syntactic features for DS-ND classification. We found all differences to be significant with a p-value of ≤ 0.001. We discuss key differences below.

Profile Differences Similar to previous work, we found that across all languages DS accounts use shorter names$\downarrow$ (Ferrara et al. 2016), have a lower follower-to-friend ratio$\downarrow$ (Lee, Eoff, and Caverlee 2011), produce less tweets (Lin and Huang 2013), and do not live long (e.g., have been active for less days) (Lee, Eoff, and Caverlee 2011; Thomas et al. 2011; Ferrara et al. 2016).

In contrast to previous work, we observed that DS accounts produce shorter bio field descriptions$\downarrow$ across all languages except for Spanish and have significantly less favorites$\downarrow$, followers$\downarrow$, and friends$\downarrow$ (except for Russian) (Lin and Huang 2013; Lee, Eoff, and Caverlee 2011; Thomas et al. 2011). It may suggest that previous findings on following and friending strategies for spam accounts is different from deleted or suspended accounts. Alternatively, content polluters may change this behavior over time. For instance, fraudulent accounts labeled as "trolls" according to (Boffey 2016) are created to look like real users. Trolls have similar follower and friend counts as the legitimate users, engage in conversations with other users, express opinions and emotions and share images.

Syntactic and Stylistic Differences Similar to previous work, we found that deleted and suspended accounts use less hashtags$\downarrow$ and mentions$\downarrow$ (except for English) (Ferrara et al. 2016; Guo and Chen 2014). In addition, we observed novel, previously unseen differences in shallow features – across all languages DS accounts use less punctuation$\downarrow$ (except for



(a) Profile differences

(b) Syntactic and sylistic differences

Figure 2: Mean differences in profile, syntactic and stylistic features between DS and ND accounts.

Spanish), repeated punctuation↓ e.g., *?????, !!!!*, capitalized words↓ e.g., *WOW*, and elongations↓ e.g., *noooo* (except for English). In contrast to previous work, we observed that deleted and suspended accounts produce less retweets↓ and URLs↓ (Ferrara et al. 2016; Thomas et al. 2011) and more emoticons (Guo and Chen 2014).

Sentiment and Emotion Differences In Figure 3 we show that on average DS accounts produce more opinionated content (less neutral) – positive and negative tweets (except for English). Previous work on applying sentiment features for influence bot detection (Dickerson, Kagan, and Subrahmanian 2014; Subrahmanian et al. 2016) observed similar behavior for English. However, our results demonstrate that these findings are not consistent across languages. We observed that DS accounts produce less anger↓ and fear↓ but more disgust↑ across all languages, and more sadness↑, surprise↑ and joy↑ (except for Spanish).

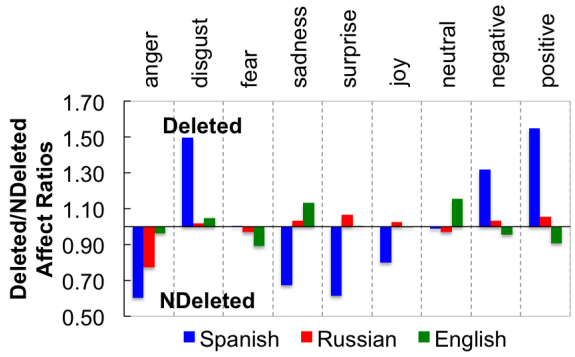


Figure 3: Mean differences in affect features between DS and ND accounts.

Summary and Discussion

Early elimination of suspicious accounts on Twitter that can potentially be spreading disinformation, deceptive and abusive content will not only reduce sampling biases when building social media analytics e.g., flu detector or personality analyzer, but is also important to ensure safer environment in social media.

We presented an approach and performed an extensive set of experiments for detecting "to be deleted or suspended" accounts on Twitter. We analyzed the predictive power of under-explored image and affect features, and text features such as topics and embeddings contrasting them with widely used network and profile signals. We have not only demonstrated that text features outperform profile and network features but also found that the presence of certain topics, hashtags, and ngrams in user tweets leads to a higher likelihood for that users' account deletion or suspension.

In contrast to previous work, we uncovered novel differences in deleted and suspended behavior of users speaking different languages. For example, we found that compared to active users deleted accounts:

- have shorted biographies in English and Russian, but not in Spanish; have less followers and friends in English and Spanish but not in Russian.
- use less hashtags and mentions, repeated punctuation, capitalizations and elongations in Russian and Spanish but not in English.
- produce more opinionated content (less neutral) – more positive and negative tweets in Spanish and Russian but not in English; more sadness, surprise and joy in English and Russian but not in Spanish.

Finally, we demonstrated that neural network models

trained on text and network features yield the highest prediction performance for the majority of classification tasks across languages. However, we found that the predictive power of different feature types is not consistent across three evaluated languages. In the future we plan to further explore these differences, and evaluate image features in combination with text features for account deletion and suspension prediction.

Acknowledgements

The authors would like to thank Josh Harrison and David Gillen for their contribution to this project, and anonymous reviewers for their helpful comments.

References

Almuhimedi, H.; Wilson, S.; Liu, B.; Sadeh, N.; and Acquisti, A. 2013. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of CSCW*, 897–908.

Atefeh, F., and Khreich, W. 2015. A survey of techniques for event detection in Twitter. *Computational Intelligence* 31(1):132–164.

Bamman, D.; O'Connor, B.; and Smith, N. 2012. Censorship and deletion practices in Chinese social media. *First Monday* 17(3).

Bamman, D. 2016. Interpretability in human-centered data science. *Workshop on Human-Centered Data Science*.

Benevenuto, F.; Magno, G.; Rodrigues, T.; and Almeida, V. 2010. Detecting spammers on Twitter. In *Proceedings of CEAS*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Boffey, D. 2016. Europes new cold war turns digital as Vladimir Putin expands media offensive.

Cao, Q.; Sirivianos, M.; Yang, X.; and Pregueiro, T. 2012. Aiding the detection of fake accounts in large scale social online services. In *9th USENIX Symposium on NSDM*.

Chetviorkin, I.; Moscow, L. G.; and Loukachevitch, N. 2014. Two-step model for sentiment lexicon extraction from Twitter streams. In *Proceedings of ACL*.

Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing* 9(6):811–824.

Dickerson, J. P.; Kagan, V.; and Subrahmanian, V. 2014. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? In *Proceedings of ASONAM*.

Dumais, S. T. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology* 38(1):188–230.

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2014. Diffusion of lexical change in social media. *PloS one* 9(11):e113114.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM* 59(7):96–104.

Ferrara, E. 2015. Manipulation and abuse on social media. *arXiv preprint arXiv:1503.03752*.

Golbeck, J.; Robles, C.; Edmondson, M.; and Turner, K. 2011. Predicting personality from Twitter. In *Proceedings of SocialCom/PASSAT*.

Guo, D., and Chen, C. 2014. Detecting non-personal and spam users on geo-tagged Twitter network. *Transactions in GIS* 18(3):370–384.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Hwang, T.; Pearce, I.; and Nanis, M. 2012. Socialbots: Voices from the fronts. *Interactions* 19(2):38–45.

Kim, S.; Weber, I.; Wei, L.; and Oh, A. 2014. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of Hypertext and Social Media*, 243–248.

Kim, J. W.; Kim, D.; Keegan, B.; Kim, J. H.; Kim, S.; and Oh, A. 2015. Social media dynamics of global co-presence during the 2014 fifa world cup. In *Proceedings of HFCS*, 2623–2632.

Lampos, V.; Aletras, N.; Preoţiuc-Pietro, D.; and Cohn, T. 2014. Predicting and characterizing user impact on Twitter. In *Proceedings of EACL*.

Lawrence, A. 2015. Social network analysis reveals full scale of Kremlin's Twitter bot campaign.

Lee, K.; Caverlee, J.; and Webb, S. 2010. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of ACM SIGIR*.

Lee, K.; Eoff, B. D.; and Caverlee, J. 2011. Seven months with the devils: A long-term study of content polluters on Twitter. In *Proceedings of ICWSM*.

Lei, T.; Zhang, Y.; Barzilay, R.; and Jaakkola, T. 2014. Low-rank tensors for scoring dependency structures. In *Proceedings of ACL*.

Lin, P.-C., and Huang, P.-M. 2013. A study of effective features for detecting long-surviving twitter spam accounts. In *Proceedings of ICACT*, 841–846.

Martinez-Romo, J., and Araujo, L. 2013. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications* 40(8).

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of ICWSM*, 122–129.

Passos, A.; Kumar, V.; and McCallum, A. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of CoNLL*.

Paul, and Dredze, M. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. *Artificial Intelligence*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, volume 14, 1532–1543.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2013. I wish i didn't say that! analyzing and predicting deleted messages in Twitter. *arXiv preprint arXiv:1305.3107*.

Potash, P.; Bell, E.; and Harrison, J. 2016. Using topic modeling and text embeddings to predict deleted tweets. *Proceedings of AAAI WIT-EC*.

Preoţiuc-Pietro, D.; Volkova, S.; Lampos, V.; Bachrach, Y.; and Aletras, N. 2015. Studying user income through language, behaviour and affect in social media. *PloS one* 10(9):e0138717.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *Computer Vision* 115(3):211–252.

Santos, C. N., and Zadrozny, B. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings ICML*.

Subrahmanian, V.; Azaria, A.; Durst, S.; Kagan, V.; Galstyan, A.; Lerman, K.; Zhu, L.; Ferrara, E.; Flammini, A.; Menczer, F.; et al. 2016. The DARPA Twitter Bot Challenge. *arXiv preprint arXiv:1601.05140*.

Thomas, K.; Grier, C.; Song, D.; and Paxson, V. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of ACM SIGCOMM*, 243–258.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*.

Volkova, S., and Bachrach, Y. 2015. On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychology, Behavior, and Social Networking* 18(12):726–736.

Volkova, S.; Chetviorkin, I.; Arendt, D.; and Van Durme, B. 2016. Contrasting public opinion dynamics and emotional response during crisis. In *Proceedings of SocInfo*.

Volkova, S.; Wilson, T.; and Yarowsky, D. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP*, 1815–1827.

Wallach, H. 2014. Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency.

Wang, G.; Mohanlal, M.; Wilson, C.; Wang, X.; Metzger, M.; Zheng, H.; and Zhao, B. Y. 2012. Social turing tests: Crowdsourcing sybil detection. *arXiv preprint arXiv:1205.3856*.

Yang, C.; Harkreader, R.; Zhang, J.; Shin, S.; and Gu, G. 2012. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, 71–80. ACM.

Yang, C.; Harkreader, R. C.; and Gu, G. 2011. Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers. In *Intrusion Detection*.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? In *Proceedings of NIPS*, 3320–3328.

Zhang, Y.; Wang, S.; Phillips, P.; and Ji, G. 2014. Binary pso with mutation operator for feature selection using decision tree applied to spam detection. *Knowledge-Based Systems* 64:22–31.