To appear in the 2014 ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media

User Type Classification of Tweets with Implications for Event Recognition

Lalindra De Silva and **Ellen Riloff**
School of Computing
University of Utah
Salt Lake City, UT 84112
{alnds,riloff}@cs.utah.edu

Abstract

Twitter has become one of the foremost platforms for information sharing. Consequently, it is beneficial for the consumers of Twitter to know the origin of a tweet, as it affects how they view and interpret this information. In this paper, we classify tweets based on their origin, *exploiting only the textual content of tweets*. Specifically, using a rich, linguistic feature set and a supervised classifier framework, we classify tweets into two user types - *organizations* and *individual persons*. Our user type classifier achieves an 89% F_1-score for identifying tweets that originate from organizations in English and an 87% F_1-score for Spanish. We also demonstrate that classifying the user type of a tweet can improve downstream event recognition tasks. We analyze several schemes that exploit user type information to enhance Twitter event recognition and show that substantial improvements can be achieved by training separate models for different user types.

1 Introduction

Twitter has become one of the most widely used social media platforms, with users (as of March 2013) posting approximately 400 million tweets per day (Wickre, 2013). This public data serves as a potential source for a multitude of information needs, but the sheer volume of tweets is a bottleneck in identifying relevant content (Becker et al., 2011). De Choudhury et al. (2012) showed that the user type of a Twitter account is an important indicator in sifting through Twitter data. The knowledge of a tweet's origin has potential implications on the nature of the content to an end user (e.g., credibility, location, etc). Also, certain types of events are more likely to be reported by individual persons (e.g., local events) whereas organizations generally report events that are of interest to a wider audience.

The first part of our research focuses on user type classification in Twitter. De Choudhury et al. (2012) addressed this problem by examining meta-information derived from the Twitter API. In contrast, the goal of our work is to classify tweets, ***based solely on their textual content***. We highlight several reasons why this can be advantageous. One reason is that people frequently share content from other sources, but the shared content often appears in their Twitter timeline as if it was their own. Consequently, a tweet that was posted by an individual may have originated from an organization. Moreover, meta-information can sometimes be infeasible to obtain given the rate limits[1] and there are times when profile information for a user account is unavailable or ambiguous (e.g., users often leave their profile information blank or write vague entries). Therefore, we believe there is value in being able to infer the type of user who authored a tweet based solely on its textual content. Potentially, our methods for user type classification based on textual content can also be combined with methods that examine user profile data or other meta-data, since they are complementary sources of information.

In this paper, we present a classifier that tries to determine whether a tweet originated from an organization or a person using a rich, linguistically-motivated feature set. We design features to recognize linguistic characteristics, including sentiment and emotion expressions, informal language use, tweet style, and similarity with news headlines. We evaluate our classifier on both English and Spanish Twitter data and find that the classifier achieves an 89% F_1-score for identifying tweets that originate from organizations in English and a

[1] https://dev.twitter.com/docs/rate-limiting/1.1/limits

87% F_1-score for Spanish.

The second contribution of this paper is to demonstrate that user type classification can improve event recognition in Twitter. We conduct a study of event recognition for civil unrest events and disease outbreak events. Based on statistics from manually annotated tweets, we found that organization-tweets are much more likely to mention these events than person-tweets. We then investigate several approaches to incorporate user type information into event recognition models. Our best results are produced by training separate event classifiers for tweets from different user types. We show that user type information consistently improves event recognition performance for both civil unrest events and disease outbreak events and for both English and Spanish tweets.

2 Related Work

Our work is most closely related to that of De Choudhury et al. (2012), which proposed methods to classify Twitter users into three categories: 1) Journalists/media bloggers, 2) Organizations and 3) Ordinary Individuals. They employed features derived from social network structure, user activity and users' social interaction behaviors, and named entities and historical topic distributions in tweets. In contrast, our work classifies isolated tweets into two different user types, based on their textual content. Consequently, our work can produce different user type labels for different tweets by the same user, which can help identify shared content not authored by the user.

Another body of related work tries to classify Twitter users along other dimensions such as ethnicity and political orientation (Pennacchiotti and Popescu, 2011; Cohen and Ruths, 2013). Gender inference in Twitter has also garnered interest in the recent past (Ciot et al., 2013; Liu and Ruths, 2013; Fink et al., 2012). Researchers have also focused on user behaviors showcased in Twitter including the types of messages posted (Naaman et al., 2010), social connections (Wu et al., 2011), user responses to events (Popescu and Pennacchiotti, 2011) and behaviors related to demographics (Volkova et al., 2013; Mislove et al., 2011; Rao et al., 2010).

Event recognition is another area that continues to attract a lot of interest in social media. Previous work has investigated event identification and extraction (Jackoway et al., 2011; Becker et al., 2009; Becker et al., 2010; Ritter et al., 2012), event discovery (Benson et al., 2011; Sakaki et al., 2010; Petrović et al., 2010), tracking events over time (Kim et al., 2012; Sayyadi et al., 2009) and event retrieval over archived Twitter data (Metzler et al., 2012). While our work focuses on user type classification, we show that the user type of a tweet is an important piece of information that can be beneficial in event recognition models.

3 Twitter User Types

Twitter user types can be analyzed in different granularities and across different dimensions. We follow a high-level categorization of user types into organizations and individual persons. While we acknowledge the existence of other user types, such as automated bots, we focus only on the *organization* and *individual person* user types for our research.

Sample Tweets from Organizations

- *Banking Commission Split Over EU Bonus Cap http://t.co/GSSbmHAWsQ*
- *Apple likely to introduce smaller, cheaper iPad mini today http://t.co/TuKBHZ3z*
- *Diet Coke may be the new #2, but U.S. soda market is shrinking http://ow.ly/1bSNnh*

Sample Tweets from Individual Persons

- *@john It's a stress fracture. Nah, no dancing was involved!*
- *My gawd feels like my head's gonna explode*
- *Watching The Rainmaker. It has totally sucked me in :D #notsomuch lol*

Figure 1: Sample tweets from individual persons and organizations

From a linguistic point of view, we can observe several distinguishing characteristics between organization- and person-tweets. As shown in Figure 1, organization-tweets are often characterized by headline-like language usage, structured style, a lack of conversation with the audience (i.e., few reply-tweets), and hyperlinks to original articles. In contrast, person-tweets show significant language variability including shorthand terms, conversational behavior, slang and profanity, expressions of emotion, and an overall relaxed usage of language.

3.1 Data Acquisition for User Types

To create our data sets, we archived tweets (using Twitter Streaming API) for six months, beginning February 1^{st}, 2013. We then used a language filter (Lui and Baldwin, 2012) to separate out the English and Spanish tweets. Also, in the data sets we created (see below), we removed duplicates, retweets and any tweet with less than 5 words. Given that large-scale human annotation is expensive, we explored several heuristics to reliably compile a large gold standard collection of person- and organization-tweets.

3.1.1 Acquiring Person-tweets

To acquire person-tweets, we devised a *person heuristic*, focusing on the *name* and the *profile description* fields in each user account corresponding to a tweet. We first gathered lists of **person names** (first names and surnames), for both English and Spanish, using census data[2] and online resources[3]. We also compiled a list of common **organization terms** (e.g., *agency*, *institute*, *company*, etc) in both English and Spanish.

The *person heuristic* labels a tweet as a person-tweet if [no organization term is in the name or the profile description fields] **AND** [all the words in the name field are person names *OR* the profile description field starts with either *'I'm'* or *'I am'*][4]. To assess the accuracy of the *person heuristic*, we also performed a manual annotation task. We employed two annotators and provided them with guidelines of what constitutes an individual person's Twitter account. We defined an individual person as someone who uses Twitter in their day-to-day life to post information about his/her daily activities, update personal status messages, comment about societal issues and/or interact with close social circles. The annotators were able to see the *name*, *profile description*, *location* and *url* fields of the Twitter user account and were asked to label each account as ***individual***, ***not individual*** or ***undetermined***. To calculate annotator agreement between the two annotators, we gave them 100 Twitter accounts, corresponding to English tweets collected using the *person heuristic*. The inter-annotator agreement (IAA) was .98 (raw agreement) and .97 (G-Index score). We did not use Cohen's Kappa (κ) as it is known to underestimate agreement (known as Kappa Paradox) when one category dominates. We then released another 250 accounts to each of the annotators, giving us a total of 600 manually labeled accounts[5].

In the distribution of labels assigned by the human annotators for these 600 accounts, 91.5% was confirmed as belonging to ***individual*** persons. 5% was identified as ***not individual*** whereas 3.5% was labeled as ***undetermined***. These numbers corroborate our claim that the *person heuristic* is a valid approximation for acquiring person-tweets.

However, limiting our person-tweets to those from accounts identified with the *person heuristic* could introduce bias (i.e., it may consider only the people who provided more complete profile information). To address this issue, we looked into additional heuristics that are representative of individual persons' Twitter accounts. We observed that applications designed specifically for hand-held devices (e.g., *twitter for iphone*) are frequently used to author tweets and used by individual persons. Organizations, on the other hand, primarily use the Twitter web tool and content management software applications to create, manage and post content to Twitter.

To further investigate our observation, we extracted the source information (i.e., the software applications used to author tweets) for a collection of 1.2 million English tweets from our tweet pool, for a random day, and identified those that were clearly *hand-held device apps* and covered at least 1% of the tweets. Table 1 shows the distribution of these *hand-held device apps*, which together accounted for approximately 66% of all tweets.

Hand-held Device App	% of Tweets
twitter for iphone	37.11
twitter for android	16.50
twitter for blackberry	5.50
twitter for ipad	2.55
mobile web (m2)	1.46
ios	1.36
echofon	1.29
ALL	**65.77**

Table 1: Percentage of (English) tweets authored from hand-held device apps

To evaluate our hypothesis that a high percentage of these tweets are person-tweets, we carried out another manual annotation task. We selected

[2] http://www.census.gov/genealogy/www/data/1990surnames/names_files.html

[3] http://genealogy.familyeducation.com/browse/origin/spanish

[4] Corresponding terms were used for Spanish

[5] We adjudicated the disagreements in the initial 100 Twitter accounts.

100 English Twitter accounts whose tweets were generated using one of the above *hand-held device apps* and asked the two annotators to label them using the same guidelines. For this task, the IAA was .84 (raw agreement) and .76 (G-Index score). As before, we released another 250 accounts to each of the annotators. In these 600 user accounts, 87.1% was confirmed to be ***individual*** persons. Only 1% was judged to be clearly ***not individual*** whereas 11.9% was labeled as ***undetermined***.

3.1.2 Acquiring Organization-tweets

Designing similar heuristics to identify organization-tweets proved to be difficult. Organizations describe themselves in numerous ways, making it difficult to automatically identify their names in user profiles. Furthermore, organization names often appear in individual persons' accounts because they mention their employers (e.g., *I'm a software engineer at Microsoft Corporation*). Therefore, to acquire organization-tweets, we relied on web-based directories of organizations (e.g., `www.twellow.com`) and gathered their tweets using the Twitter API. We used 58 organization accounts for English tweets and 83 accounts for Spanish.

3.1.3 Complete Data Set

We created a data set of 200,000 tweets for each language, consisting of 90% person-tweets and 10% organization-tweets. Among the 180,000 person-tweets, 132,000 (66% of 200,000) were tweets whose source was a *hand-held device app*. To collect the remaining 48,000 (24% of 200,000) of the person-tweets, we relied on the *person heuristic*. Finally, we gathered 20,000 organization-tweets using the web directories mentioned previously. In doing so, to ensure that we had a balanced mix of organizations, each organization contributed with a maximum of 500 tweets.

4 User Type Classification

To automatically distinguish person-tweets from organization-tweets, we trained a supervised classifier using N-gram features, an organization heuristic, and a linguistic feature set categorized into six classes. For the classification algorithm, we employed a Support Vector Machine (SVM) with a linear kernel, using the LIBSVM package (Chang and Lin, 2011). For the features that rely on part-of-speech (POS) tags, we used the English Twitter POS tagger by Gimpel et al. (2011) and another tagger trained on the CoNLL 2002 shared task data for Spanish (Tjong Kim Sang, 2002) using the OpenNLP toolkit (OpenSource, 2010).

4.1 N-gram Features

We started off by introducing N-gram features to capture the words in a tweet. Specifically, we trained a supervised classifier using unigram and bigram features encoded with binary values. In selecting the N-gram features, we discarded any N-gram that appears less than five times in the training data.

4.2 Organization Heuristic

Following observations by Messner et al. (2011), we combined two simple heuristic rules to flag tweets that are likely to be from an organization. The first observation is that 'replies' (i.e., @user mentions at the beginning of a tweet) are uncommon in organization-tweets. Hence, if a tweet is a reply, it is likely to be a person-tweet. The second observation is that organization-tweets frequently include a web link to external content.

Our ***organization heuristic***, therefore, combined these two properties. If the tweet is not a reply **AND** contains a web link, we labeled it as an organization-tweet. Otherwise, we labeled it as a person-tweet. In Section 5, we evaluate this heuristic as a classification rule on its own, and also incorporate its label as a feature in our classifier.

4.3 Linguistic Features

In the following sections, we describe our linguistic features and the intuitions in designing them.

4.3.1 Emotion and Sentiment

Twitter is a platform where individuals often express emotion. We detected emotions using four feature types: 1) interjections, 2) profanity, 3) emoticons and 4) overall sentiment of the tweet.

Interjections, profanity, and emoticons are widely used by individuals to convey emotion, such as anger, surprise, happiness, etc. To identify these three feature types, we used a combination of POS tags in the English tagger (which contains tags for interjections, emoticons, etc), compiled lists of interjections and profanity from the

web for both English[6] and Spanish[7] and regular expression patterns for emoticons.

We also included sentiment features using the sentiment140 API[8] (Go et al., 2009). This API provides a sentiment label (positive, negative or neutral) for a tweet corresponding to its overall sentiment. We expect person-tweets to show more positive and negative sentiment and organization-tweets to be more neutral.

4.3.2 Similarity to News Headlines

Earlier, we observed that organization-tweets are often similar to news headlines. To exploit this observation, we introduced four features using language models and verb categories.

First, we collected 3 million person-tweets, for each language, using the *person heuristic* described in Section 3.1. Second, we collected another 3 million news headlines from each of the English and Spanish Gigaword corpora (Parker et al., 2009; Mendonca et al., 2009). Using these two data sets, we built unigram and bigram language models (with Laplace smoothing) for person-tweets and for news headlines. Given a new tweet, we calculated the probability of the tweet using both the person-tweet and headline language models. We defined a binary feature that indicates which unigram language model (person-tweet model *vs.* headline model) produced the highest probability. A similar feature is defined for the bigram language models.

We also observed that certain verbs are predominantly used in news headlines and are rarely associated with colloquial language (therefore, in person-tweets). Similarly, we observed verbs that are much more likely to be used by individual persons. To identify the most discriminating verbs, we ranked verbs appearing more than 5 times in the collected news headlines and person-tweets based on the following probabilities:

$$p(h|verb) = \frac{\text{Frequency of } verb \text{ in } \textit{headlines}}{\text{Frequency of } verb \text{ in } \textit{all instances}}$$

$$p(pt|verb) = \frac{\text{Frequency of } verb \text{ in } \textit{person-tweets}}{\text{Frequency of } verb \text{ in } \textit{all instances}}$$

The verbs were sorted by probability and we retained two disjoint sets of verbs, 1) the verbs most representative of headlines (i.e., *headline verbs*), selected by applying a threshold of $p(h|verb) > 0.8$ and 2) verbs most representative of person-tweets (i.e., *personal verbs*), with a similar threshold of $p(pt|verb) > 0.8$. We introduced two binary features that look for verbs in the tweet from these two learned verb lists. The top-ranked verbs for each category are displayed in Table 2. The learned headline verbs tend to be more formal and are often used in business or government contexts (e.g., *revoke*, *granting*, *etc*) whereas the personal verbs tend to represent personal activities, communications, and emotions (e.g., *hate*, *sleep*, *etc*). In total, we learned 687 headline verbs and 2221 personal verbs for English, and 1924 headline verbs and 5719 personal verbs for Spanish.

Headline verbs: aided, revoke, issued, broaden, testify, leads, postponing, forged, deepen, hijacked, raises, granting, honoring, pledged, departing, suspending, citing, compensate, preserved, weakening, differing
Personal verbs: raining, sleep, hanging, hate, marching, teaching, sway, having, risk, lurk, screaming, tagging, disturb, baking, exaggerate, pinch, enjoy, shredding, force, hide, wreck, saved, cooking, blur, told

Table 2: Top-ranked representative verbs learned from headlines and person-tweets

4.3.3 1st and 2nd Person Pronouns

Person-tweets often include self-references, in the form of first-person pronouns and their variant forms (e.g., possessive, reflexive), while organization-tweets rarely contain self-references. Also, organizations often address their audience using second-person pronouns in tweets (e.g., *Will **you** High Five the #Bruins or #Blackhawks? Sign up for a chance to win a trip to the Cup Final: http://t.co/XQP8ZDOINV*). To capture these characteristics, we included two binary features that look for 1st and 2nd person pronouns in a tweet.

4.3.4 NER Features

We hypothesized that organization-tweets will carry more named entities and proper nouns. For English tweets, we identified *Persons*, *Organizations* and *Locations* using the Named Entity Recognizer (NER) from Ritter et al. (2011). For Spanish tweets, we used NER models trained on CoNLL 2002 shared task data for Spanish. The features were encoded as three values, representing the frequency of each entity type in a tweet.

[6] http://www.noswearing.com/dictionary

[7] http://nawcom.com/swearing/mexican_spanish.htm

[8] http://help.sentiment140.com/api

	English			Spanish		
	P	**R**	**F_1**	**P**	**R**	**F_1**
ULM: Unigram Language Model	71.63	63.18	67.14	66.14	60.43	63.16
BLM: Bigram Language Model	81.46	49.17	61.32	80.03	51.08	62.36
NGR: SVM with N-grams	86.02	62.76	72.57	85.76	66.56	74.95
OrgH: Organization Heuristic	66.87	**91.08**	77.12	65.32	81.44	72.49
NGR + OrgH	82.26	86.82	84.48	83.85	85.17	84.50
NGR + OrgH + Linguistic Features	**89.01**	89.40	**89.20**†	**87.59**	**85.47**	**86.52**†

Table 3: User type classification results with **P**recision (%), **R**ecall (%) and **F_1**-Score (%). † denotes statistical significance at $p < 0.01$ compared to *NGR + OrgH*

4.3.5 Informal Language Features

Person-tweets often showcase erratic and casual use of language, whereas organization-tweets tend to have (relatively) more grammatical language usage. Hence, we introduced a feature to determine the *informality* of a tweet. Specifically, we check if a tweet begins with an uppercase letter or not, and whether sentences are properly separated with punctuation. To accomplish this, we used regular expression patterns that look for capitalized characters following punctuation and white-space characters. We also added a feature to check if all the letters in the tweet are lowercased. Use of elongated words (e.g., *cooooooool*) for emphasis, is another property of person-tweets and we captured this property by identifying words with three or more repetitions of the same character.

To comply with the 140 character length restriction of a tweet, person-tweets often employ ad-hoc short-hand usage of words that omit or replace characters with a phonetic substitute (e.g., *2mrw*, *good n8*). We used lists of common abbreviations found in social media[9] collected from the web and a binary feature was set if a tweet contained an instance from these lists.

4.3.6 Twitter Stylistic Features

One can also notice structural properties that are prevalent in either user type. News organizations often append a topic descriptor to the beginning of a tweet (e.g., ***Petraeus affair*: *Woman who complained of harassing emails identified http://t.co/hpyLQYeL***). To encode this behavior, we employed a simple heuristic that looked for a semicolon or a hyphen within the first three words of a tweet. Also, person-tweets employ heavy use of hashtags so we included the frequency of hashtags in a tweet as a single feature. We added two more features in the form of the length of the tweet and the frequency of @user mentions in the tweet.

[9] `http://www.noslang.com/dictionary/full/`

5 Evaluation of User Type Classification

In this section, we discuss and evaluate our user type classifier. All of the experiments were carried out using five-fold cross-validation, using data sets described in Section 3.1. In these experiments, we maintained the separation of organization-tweets at a user-account level in order to avoid tweets from one organization appearing in both train and test sets.

5.1 User Type Classifier Results

We first evaluated several baseline systems to assess the difficulty of the user type classification task. We report precision, recall and F_1-score with organization-tweets as the positive class.

To evaluate our hypothesis that organization-tweets are similar to news headlines, we first predicted user types using only the unigram and bigram language models described in Section 4.3.2. As shown in Table 3 (***ULM*** & ***BLM***), unigram models were capable of discerning organization-tweets with 71% and 66% precision on English and Spanish tweets, respectively. This is substantial performance given that the random chance of labeling an organization-tweet (i.e., precision) is merely 10%. The bigram models show $\geq$ 80% precision whereas the unigram models show higher recall.

As another baseline, we evaluated an SVM classifier that uses only N-gram features. As Table 3 shows, the N-gram classifier (***NGR***) achieved very high precision (86%) for both English and Spanish tweets. However, it yielded relatively moderate recall (63% for English and 67% for Spanish).

We then evaluated the organization heuristic (***OrgH***) all by itself. The heuristic identifies two common characteristics of organization-tweets and as expected, it achieved substantial recall (91% for English and 81% for Spanish) but

	English			Spanish		
	P	**R**	**F_1**	**P**	**R**	**F_1**
NGR + OrgH	82.26	86.82	84.48	83.85	85.17	84.50
+ Emotion and Sentiment Features	86.58	86.41	86.50	85.91	84.19	85.05
+ Features Derived from News Headlines	87.83	87.10	87.46	86.68	84.05	85.35
+ 1st and 2nd Person Pronouns	87.88	88.53	88.20	86.61	84.38	85.48
+ NER Features	88.05	88.75	88.40	86.71	84.69	85.69
+ Informal Language Features	88.39	89.14	88.77	86.89	85.31	86.09
+ Twitter Stylistic Features	89.01	89.40	89.20	87.59	85.47	86.52
NGR + OrgH + Linguistic Features	**89.01**	**89.40**	**89.20**	**87.59**	**85.47**	**86.52**

Table 4: Linguistic feature ablation with **P**recision (%), **R**ecall (%) and **F_1**-Score (%)

with mediocre precision.

These results show that the N-gram classifier achieved high precision whereas the organization heuristic achieved high recall. To exploit the best of both worlds, we evaluated another model (***NGR + OrgH***) that added the organization heuristic as an additional feature for the N-gram classifier. This system fares better than all the previous models, achieving 82% precision with 87% recall for English and 84% precision with 85% recall for Spanish.

Next, we show the benefits obtained from adding the linguistic feature set. As the final row in Table 3 shows, having incorporated all the linguistic features, our final system showed an improvement of 7% precision and 3% recall on English tweets for an overall F_1-score gain of approximately 5%. On Spanish tweets, the same increments were 4%, 0.3% and 2%, respectively. This final classifier is statistically significantly better than the model without linguistic features (***NGR + OrgH***) for both languages at the $p < 0.01$ level, analyzed using a paired booststrap test drawing 10^6 samples with repetition from test data, as described in Berg-Kirkpatrick et al. (2012).

5.2 Analysis of Linguistic Features

Having observed that linguistic features improved user type classification, we evaluated the impact of each type of linguistic feature using an ablation study. Table 4 shows the classifier performance when each of the features types was added cumulatively over the ***NGR + OrgH*** baseline.

We immediately see a 4% and 2% precision gain by adding emotion and sentiment features, for English and Spanish, respectively. Adding features derived from news headlines, we observe that the classifier fares better, improving precision for both languages and improving recall for English. 1st and 2nd person pronouns improved recall on English data but had little impact on Spanish data. The NER features produced very small gains in both languages. The informal language features increased recall from 84.69% to 85.31% on Spanish tweets. Finally, the Twitter stylistic features gained 0.7% more precision for both languages. Overall, the feature types that contributed the most were the emotion/sentiment features, the news headline features, and the Twitter stylistic features.

6 Twitter Event Recognition

Twitter provides a facility where users can search for tweets using keywords. However, keyword-based queries for events can often lead to myriad irrelevant results due to different senses of keywords (polysemy) and figurative or metaphorical use of keywords. For instance, a Twitter search for civil unrest events with a few representative keywords (e.g., *strike*, *rally*, *riot*, *etc.*) can often lead to results referring to sports events, such as a *bowling strike* or a *tennis rally* or where the keywords are used figuratively (e.g., *She's a riot!*). In this section, we investigate if the user type of a tweet can help cut through such ambiguity. Specifically, we hypothesize that event keywords may be used more consistantly and with less ambiguity in organization-tweets, and therefore user type information may be helpful in improving event recognition.

To explore our hypothesis that the user type can influence the event relevance of a tweet, we constructed a set of experiments using two types of events - civil unrest events and disease outbreaks. Civil unrest refers to forms of public disturbance that affect the order of a society (e.g., *strikes*, *protests*, *etc.*) whereas a disease outbreak refers to an unusual or widespread occurrence of a disease (e.g., *a measles outbreak*).

	English		Spanish	
	Civil Unrest	**Disease Outbreaks**	**Civil Unrest**	**Disease Outbreaks**
Person-tweets	5.27%	9.52%	9.32%	5.00%
Organization-tweets	36.54%	39.34%	51.66%	44.06%
All-tweets	**12.50%**	**20.07%**	**14.72%**	**13.22%**

Table 6: Percentage of event-relevant tweets in 4000 tweets with keywords for each category

English Civil Unrest: protest, protested, protesting, riot, rioted, rioting, rally, rallied, rallying, marched, marching, strike, striked, striking
English Disease Outbreaks: outbreak, epidemic, influenza, h1n1, h5n1, pandemic, quarantine, cholera, ebola, flu, malaria, dengue, hepatitis, measles
Spanish Civil Unrest: protesta, protestar, amotinaron, protestaron, protestaban, protestado, amotinarse, amotinaban, marcha, huelga, amotinando, protestando, amotinado
Spanish Disease Outbreaks: brote, epidemia, influenza, h1n1, h5n1, pandemia, cuarantena, sarampión, cólera, ebola, malaria, dengue, hepatitis, gripe

Table 5: Keywords used to query Twitter for two types of events in English and Spanish

6.1 Data Acquisition for Event Recognition

We began by collecting tweets that *contained at least one of the keywords* listed in Table 5, using the Twitter search API, and we set up an annotation task using Amazon Mechanical Turk (AMT) annotators. First, we created guidelines to distinguish event-relevant tweets from irrelevant tweets and annotated 300 tweets for each of the four categories (i.e., English Civil Unrest, Spanish Civil Unrest, English Disease Outbreaks and Spanish Disease Outbreaks).

We released 200 tweets in each category for annotation to three AMT annotators[10]. We used these 200 tweets to calculate pair-wise IAA using Cohen's Kappa (κ) which we report in Table 7. The IAA scores were generally good, ranging from 0.67 to 0.89. Each annotator subsequently labeled 2000 tweets, yielding a total of 6000 tweets for each category. In each of these 6000 tweet sets, we randomly separated 2000 tweets as tuning data and 4000 as test data.

First, we applied our user type classifier to these tweets and analyzed the number of true event tweets for each user type. Table 6 shows the percentage of true event tweets in the entire test set, as well as the percentage of event tweets for each user type. Overall, the percentage of true event tweets in each test set is $\leq$ 20%. This means that most of the tweets ($>$ 80%) with event keywords ***do not*** discuss an event, confirming the unreliability of using event keywords alone.

	English	Spanish
Civil Unrest	.89, .88, .77	.74, .74, .67
Disease Outbreaks	.82, .73, .68	.84, .83, .80

Table 7: Pair-wise inter-annotator agreement (IAA) measured using Cohen's Kappa (κ) on 200 tweets among the three AMT annotators for each event type in each language

However, there is a substantial difference in the density of true event tweets between the two user types. Across both civil unrest and disease outbreaks, and for both languages, we see a much higher percentage of organization-tweets with event keywords mentioning an event than person-tweets with event keywords. Table 6 shows that, in English civil unrest category, organization-tweets are 7 times more likely (36.54% as opposed to 5.27%) to report an actual event than person-tweets with the same keywords. In the English disease outbreaks category, organization-tweets are 4 times more likely to report an event (39.34% vs. 9.52%). We notice similar observations in the Spanish tweets too.

6.2 Event Recognition Results

In this section, we evaluate the impact of user type information by introducing a simple baseline experiment for Twitter event recognition followed by several schemes that we devised to incorporate user type information in more principled ways.

First, we trained a supervised classifier to predict the probability of a tweet being event-relevant using only unigrams and bigrams as features, encoded with binary values. This baseline system is *agnostic to the user type*. We used the SVM Platt method implementation of LIBSVM (Chang and Lin, 2011) and carried out experiments using five-fold cross-validation. As Table 8 shows, this ap-

[10]We first released 100 tweets in each category to AMT and enlisted 10 annotators. After calculating IAA on these 100 tweets, we retained 3 annotators who had the highest agreement with our annotations.

	English			Spanish		
	P	R	F_1	P	R	F_1
Civil Unrest Events						
User type-agnostic classifier	**80.97**	50.20	61.98	77.51	60.37	67.88
User type included as a feature	80.00	50.40	61.84	77.19	61.56	68.50
(θ_p, θ_o) optimized for F_1-score	60.50	**72.60**	66.00	64.97	78.57	71.13
User type-specific classifier	79.34	63.61	**70.61**†	**79.20**	**81.89**	**80.52**†
Disease Outbreak Events						
User type-agnostic classifier	83.15	55.99	66.92	80.49	56.14	66.15
User type included as a feature	**83.46**	55.36	66.57	80.93	59.36	68.48
(θ_p, θ_o) optimized for F_1-score	75.10	**66.58**	70.58	68.94	72.58	70.71
User type-specific classifier	80.35	66.07	**72.51**†	**82.20**	**74.26**	**78.03**†

Table 8: Event recognition results showing **P**recision (%), **R**ecall (%) and $\mathbf{F}_1$-Score (%), for the two event types in English and Spanish. † denotes statistical significance at $p < 0.01$ compared to the baseline (*User type-agnostic classifier*)

proach achieved 62% F_1-score in English and 68% F_1-score in Spanish, for civil unrest events. For disease outbreak events, the corresponding values were 67% and 66%.

As our first attempt to incorporate user type information, we added the user type label as one additional feature. As shown in Table 8, the added feature yielded small gains for Spanish but made little difference for English.

Given our initial hypothesis (and evidence in Table 6) about events and organization-tweets, we would prefer to be aggressive in labeling organization-tweets as event-relevant. One way to accomplish this with a trained probabilistic classifier is to assign different probability thresholds to person- and organization-tweets. To acquire the optimal threshold parameters for person-tweets (θ_p) and organization-tweets (θ_o), we performed a grid-based threshold sweep on tuning data and optimized with respect to F_1-scores. Table 8 shows that this approach yielded substantial recall gains for all four categories and produced the best F_1-scores thus far.

A more principled approach is to create two completely different classifiers, one for each user type. Each classifier can then model the vocabulary and word associations that are most likely to occur in tweets of that type. Using five-fold cross-validation, we train separate models for person- and organization-tweets. During event recognition, we first apply our user type classifier to a tweet and then apply the appropriate event recognition model. As shown in the final rows in Table 8, this method consistently outperforms the other approaches. Compared to the best competing method, the user type-specific classifiers produced F_1-score gains of 4.6% and 9.4% for English and Spanish civil unrest events, and F_1-score gains of 2% and 7.3% for English and Spanish disease outbreak events.

7 Conclusion

In this work, we tackled the problem of classifying tweets into two user types, organizations and individual persons, based on their textual content. We designed a rich set of features that exploit different linguistic aspects of tweet content, and demonstrated that our classifier achieves F_1-scores of 89% for English and 87% for Spanish. We also presented results showing that organization-tweets with event keywords have a much higher density of event mentions than person-tweets with the same keywords and showed the benefits of incorporating user type information into event recognition models. Our results showed that creating separate event recognition classifiers for different user types yields substantially better performance than using a single event recognition model on all tweets.

8 Acknowledgments

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC00285. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Hila Becker, Mor Naaman, and Luis Gravano. 2009. Event identification in social media. In *WebDB*.

Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 291–300, New York, NY, USA. ACM.

H. Becker, M. Naaman, and L. Gravano. 2011. Selecting quality twitter content for events. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM11)*.

E. Benson, A. Haghighi, and R. Barzilay. 2011. Event discovery in social media feeds. In *The 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, USA. To appear*.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on twitter: Its not easy! In *Seventh International AAAI Conference on Weblogs and Social Media*.

M. De Choudhury, N. Diakopoulos, and M. Naaman. 2012. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 241–244. ACM.

Clayton Fink, Jonathon Kopecky, and Maksym Morawski. 2012. Inferring gender from the content of tweets: A region specific example. In *ICWSM*.

K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N.A. Smith. 2011. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

A. Jackoway, H. Samet, and J. Sankaranarayanan. 2011. Identification of live news events using twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, page 9. ACM.

M. Kim, L. Xie, and P. Christen. 2012. Event diffusion patterns in social media. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Wendy Liu and Derek Ruths. 2013. Whats in a name? using first names as features for gender inference in twitter.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30. Association for Computational Linguistics.

Angelo Mendonca, David Andrew Graff, Denise DiPersio, Linguistic Data Consortium, et al. 2009. *Spanish gigaword second edition*. Linguistic Data Consortium.

M. Messner, M. Linke, and A. Eford. 2011. Shoveling tweets: An analysis of the microblogging engagement of traditional news organizations. In *International Symposium on Online Journalism, UT Austin, available at: http://online. journalism. utexas. edu/2011/papers/Messner2011. pdf (last accessed April 3, 2011)*.

D. Metzler, C. Cai, and E. Hovy. 2012. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 646–655.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the demographics of twitter users. *ICWSM*, 11:5th.

M. Naaman, J. Boase, and C.H. Lai. 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM.

OpenSource. 2010. Opennlp: $http://opennlp.sourceforge.net/$.

Robert Parker, Linguistic Data Consortium, et al. 2009. *English gigaword fourth edition*. Linguistic Data Consortium.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics.

A.M. Popescu and M. Pennacchiotti. 2011. Dancing with the stars, nba games, politics: An exploration of twitter users response to events. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Ritter, O. Etzioni, S. Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM.

T. Sakaki, M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

H. Sayyadi, M. Hurst, and A. Maykov. 2009. Event detection and tracking in social streams. In *Proceedings of International Conference on Weblogs and Social Media (ICWSM)*.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*.

K Wickre. 2013. Celebrating #twitter7. `https://blog.twitter.com/2013/celebrating-twitter7`. Accessed: 03/20/2014.

S. Wu, J.M. Hofman, W.A. Mason, and D.J. Watts. 2011. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714. ACM.