How People use Twitter in Different Languages

Wouter Weerkamp
ISLA, University of Amsterdam
w.weerkamp@uva.nl

Simon Carter
ISLA, University of Amsterdam
s.c.carter@uva.nl

Manos Tsagkias
ISLA, University of Amsterdam
e.tsagkias@uva.nl

ABSTRACT

In this paper we describe how Twitter is used in various languages. We observe notable differences between languages regarding the use of hashtags, links, mentions, and conversations. We propose two dimensions that can be used to classify languages, each of which is likely to require different ways of analysis.

Categories and Subject Descriptors

K.4.0 [**Computing Milieux**]: Computers and Society—*General*

General Terms

Human Factors, Languages

1. INTRODUCTION

Microblogging platform Twitter[1] has become one of the most important real-time information resources [3]. Microblogging platforms in general offer a broad range of uses and applications, including event detection [5, 7], media analysis [1], and mining consumer and political opinions [4, 6].

Usage of Twitter is not just limited to the US or to the English-speaking world. Other countries, like Japan, Indonesia, Brazil, Germany, and the Netherlands actively participate on Twitter, and contribute to a large degree to what is discussed in the microblogosphere. Although identification of languages in tweets might be harder than in formal text, it is possible using Twitter specific priors [2].

We are interested in the way people use Twitter in different languages, and would like to see if there are obvious differences between languages in the usage of Twitter features. For this, we look at four Twitter specific features, *hashtags*, *links*, *mentions*, and *conversations*, and explore their usage in eight popular Twitter languages (Dutch, English, German, French, Indonesian, Japanese, Portuguese, and Spanish). We use the approach in [2] to identify the language of tweets, and construct a set of 1,000 tweets per language for our exploration.

[1] http://www.twitter.com

In the remainder of this paper we look at each of the features individually (Sections 1.1 to 1.4), and draw preliminary conclusions in Section 2.

1.1 Hashtags

Hashtags allow users to "tag" their tweet. In many cases this tag is one term long, but people also concatenate several words into one hashtag. The tags are mainly used to indicate the topic of interest in the tweet, and the hashtags allow, for example, for easy assessment of trending topics. Table 1 shows the statistics of hashtag usage in our set of languages; we report on the percentage of tweets having at least one hashtag, and on the average number of hashtags per *tagged* tweet.

Language	tagged tweets	avg. tags/ tweet
Dutch	16%	1.3
English	14%	1.4
German	25%	1.9
French	16%	1.4
Indonesian	10%	1.1
Japanese	4%	1.2
Portuguese	11%	1.5
Spanish	12%	1.3

Table 1: Hashtag usage per language.

We make four observations: (1) German tweets are much more likely to contain hashtags than any of the other languages, (2) The use of hashtags in German is popular, with one in every four tweets containing at least one hashtag, (3) The number of tags per tweet is much higher in German than in other languages, and (4) Indonesian and especially Japanese tweets are unlikely to contain hashtags, with only one in every 25 Japanese tweets containing a hashtag.

1.2 Links

As in other online content, tweets can contain links to other web pages. Since tweets are only very short (140 characters long), a tweet usually contains at most one link. In Table 2 we report on the percentage of tweets that contain a link for each language.

Here, we observe two things: (1) Adding links to tweets is very popular in German tweets, with close to 50% of tweets containing a link; (2) In Dutch, Indonesian, Japanese, and Portuguese tweets, adding links is not popular, since only 10–15% contains a link. This is considerably less than the other four languages.

Language	linked tweets
Dutch	15%
English	30%
German	48%
French	37%
Indonesian	12%
Japanese	11%
Portuguese	10%
Spanish	24%

Table 2: Link usage per language.

1.3 Mentions

A mention in a tweet is recognizable by the @ sign followed by a username, and indicates that someone aims her tweet directly at that person. Mentions are a more "social" feature than hashtags and links, and indicate personal communication between Twitter users. Table 3 shows the percentage of tweets that contain at least one mention, and the average number of mentions in tweets that have mentions.

Language	mentioned tweets	avg. mentions/tweet
Dutch	62%	1.2
English	50%	1.1
German	28%	1.1
French	55%	1.2
Indonesian	77%	1.8
Japanese	48%	1.2
Portuguese	45%	1.2
Spanish	62%	1.2

Table 3: Mention usage per language.

We see that four groups appear in our languages: (i) very high mention usage for Indonesian; (ii) high mention usage for Dutch and Spanish; (iii) medium-high usage for English, French, Japanese, and Portuguese, and (iv) low usage for German. Another interesting point is that the popularity of mentions does not necessarily influence the number of mentions per tweet (which was the case for hashtags). Although Indonesian does have 1.8 mentions on average, this number is very similar for the remaining seven languages.

1.4 Conversations

The final feature we explore are conversations: Twitter allows users to explicitly reply to other users' tweets, and thereby entering a conversation. Like the mentions feature, conversations is a more social aspect of Twitter than hashtags and links. In Table 4 we list the percentage of tweets that are part of a conversation.

For most languages we observe similar behavior for conversations as for the usage of mentions: conversations are popular among Dutch and Spanish tweets (one in every three tweets is part of a conversation), and less popular for French and English. It is interesting to see that Indonesian tweets have the lowest percentage of conversations, even though they had, by far, most mentions. Portuguese and Japanese, both very similar in usage of mentions, show a large difference on conversations, Japanese tweets being twice as often part of a conversation. Finally, we see that German tweets are also very unlikely to be part of a conversation, just as their percentage of mentions was very low.

Language	tweets in conversation
Dutch	36%
English	25%
German	14%
French	27%
Indonesian	13%
Japanese	26%
Portuguese	13%
Spanish	34%

Table 4: Conversations per language.

2. CONCLUSIONS

We explore how people use Twitter in different languages, and observe large differences in the use of Twitter specific features. We propose two dimensions that can be used to classify languages in Twitter. The first dimension is *structure* and indicates to what extent people add structure to tweets by adding hashtags and links. The second dimension is the *communication paradigm*, which indicates if people use Twitter as broadcasting channel (i.e., one-to-many) or as personal communication channel (i.e., one-to-one).

German tweets can be classified as structured broadcasts, characterized by high usage of hastags and links and a limited usage of personal communication options. Spanish and Dutch tweets on the other hand, are examples of mostly unstructured personal communications: limited usage of hashtags and links, but many mentions and conversations. These usage differences between languages calls for different analysis methods: German tweets can benefit greatly from hashtag analysis, and Dutch and Spanish tweets are more likely to benefit from, for example, social network analysis. Extending this exploration to more languages could reveal more tweet classes like the two mentioned here, and language groups that share similar usage patterns.

3. REFERENCES

[1] D. L. Altheide. *Qualitative Media Analysis (Qualitative Research Methods)*. Sage Pubn Inc, 1996.

[2] S. Carter, M. Tsagkias, and W. Weerkamp. Semi-supervised priors for microblog language identification. In *Dutch-Belgian Information Retrieval workshop (DIR 2011)*, 2011.

[3] G. Golovchinsky and M. Efron. Making sense of twitter search. In *Proceedings of CHI 2010 Workshop on Microblogging: What and How Can We Learn From It?*, 2010.

[4] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60 (11):2169–2188, 2009.

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

[6] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, pages 178–185, 2010.

[7] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems (CHI 2010)*, pages 1079–1088, 2010.