A Gender Based Study of Tagging Behavior in Twitter

Evandro Cunha[1], Gabriel Magno[1], Virgilio Almeida[1], Marcos André Gonçalves[1], Fabrício Benevenuto[2]

[1]Computer Science Department, Federal University of Minas Gerais (UFMG), Brazil
[2]Computer Science Department, Federal University of Ouro Preto (UFOP), Brazil

{evandrocunha, magno, virgilio, mgoncalv, fabricio}@dcc.ufmg.br

ABSTRACT

Gender plays a key role in the process of language variation. Men and women use language in different ways, according to the expected behavior patterns associated with their status in the communities. In this paper, we present a first description of gender distinctions in the usage of Twitter hashtags. After analyzing data collected from more than 650,000 tagged tweets concerning three different subjects, we concluded that gender can be considered a social factor that influences the user's choice of particular hashtags about a given topic. This study aims to increase knowledge about human behavior in free tagging environments and may be useful to the development of tag recommendation systems based on users' collective preferences.

Categories and Subject Descriptors

K.4.2 [**Computers and Society**]: Social Issues

General Terms

Human Factors

Keywords

Twitter, tagging, gender studies.

1. INTRODUCTION

Behavioral differences between men and women have been studied in many fields of knowledge. Knowing them can help us to better understand not only the characteristics of individuals, but also the properties of the communities in which they take part and especially the social dynamics between genders. Studying the behavior of Web and social networks users is also a major field of study in social informatics. The Web reflects interests and values of human society, operating as a mirror to which scientists may look to analyze communities across an extremely large information space. In addition, the description of the behaviors of individuals on the Web may be useful to offer more tailored services according to their characteristics, preferences and requirements. Supported by studies showing that men and women tend to deal with linguistic elements and innovations in different ways, we investigate the conduct of both genders regarding a specific the use of hashtags in Twitter.

The fact that any term can be turned into a hashtag and disseminate through the network generates interest in studying the dynamics of its creation, use and propagation, as a way to understand the behavior of users in free tagging environments. Our aim is to complement the experimental studies regarding

these dynamics in the light of linguistics and gender theory, looking for a description of the roles that men and women have in the life process of this kind of linguistic element.

The analysis of social factors that influence the choice of a linguistic variant is an important goal within sociolinguistics research. Many studies correlated gender to linguistic variation [2] and found significant differences in the way men and women speak and interact using language. These results support that the correlation between gender and linguistic variation must be associated with the social organization of the communities.

Our findings show that it is possible to recognize gendered hashtags. We identified linguistic characteristics that these elements may share and verified oppositions between different categories of tags. We noticed that transparent hashtags, which carry full and clear information about the topic, are more frequent among women. On the other hand, men seem more comfortable to use innovative (and sometimes more opaque) forms. We also show that male and female users adopt different persuasive strategies in the domain of Twitter communication, reflecting their behaviors in offline communities.

2. GENDER ANALYSIS

In this work, the main question we address is if male and female Twitter users choose the same tags when talking about the same topic. We expect to answer this question and, if there are "neutral", "male" and "female" forms, then we aim to identify some of the features that distinguish these three groups of tags.

For this study, we use three datasets to verify the behavior of both genders regarding topics of different natures. Our datasets consist of hashtags related to specific topics and their frequencies of utilization by male and female users. We selected the following topics: a) the 2010 Brazilian presidential election (dataset BE); b) the death of Michael Jackson, in 2009 (dataset MJ); and c) the swine flu outbreak of 2009 (dataset SF). Dataset BE has been built using data obtained by the National Institute of Science and Technology for the Web using a Twitter API. Datasets MJ and SF have been created from data collected by the Twitter Project at Max Planck Institute [1].

Dataset BE has also been divided into four sub-datasets. The purpose of this split is that we intend to investigate the choices of users of each gender in situations that admit linguistic variation – that is, the usage of different forms (in this case, hashtags) even when the semantic and the functional values are equivalent. Then, we created sub-datasets formed by hashtags related to the following topics: Dilma's supporters (BE-1); Serra's supporters (BE-2); Dilma's opposers (BE-3); and Serra's opposers (BE-4) (Dilma and Serra were the main candidates during 2010 Brazilian presidential election).

As our datasets include users' given names, we assigned to each user a gender based on his/her name. For doing this, we used lists

of the most common male and female names in Brazil (for dataset BE) and in the United States (for datasets MJ and SF).

For each hashtag, we calculated the percentage of occurrences generated by male and female users. As the total number of occurrences of hashtags used by men and women are different for all datasets, we converted the raw scores to the same unit of measurement using z-scores, which represent the number of standard deviation units a raw score is above or below the mean. High raw scores are above the mean and have positive z-scores; low raw scores, thus, have negative z-scores. Here, the use of z-scores does not operate as a test of statistical significance, but as a scaling factor so that comparisons between male and female users can take place using a common yard stick.

In this way, for a given hashtag, we obtain two z-scores: the "female z-score" indicates the correspondent weight of the female usage, and the "male z-score" indicates the weight of the male usage regarding that specific hashtag. "Female" and "male z-scores" are complementary, so their sum is always equal to zero. For convenience, all z-scores presented in this study will be associated to the female z-scores. Therefore, positive z-scores will indicate a prevalence of female users and negative z-scores will indicate a prevalence of male users. We found that neutral hashtags correspond to great part of the corpus, but also that there is a significant presence of gendered tags.

Inspired by linguistic and social studies, we analyzed four different opposite categories and observed the average z-scores obtained for the hashtags that belong to them. Several classic studies [2] show that, in general, women tend to use more standard and stigma-free forms than men, who usually feel more comfortable than female speakers to use substandard and sometimes innovative variants. We checked if the opposition between standard and substandard hashtags might be a factor that affects their acceptation by male and female users.

We decided to consider as "standard forms" the most used hashtags, as we noticed that they tend to be very transparent about the topic to which they refer. For instance, the most used hashtags in the datasets MJ and SF are respectively #michaeljackson and #swineflu. On the other hand, we considered the less common – and generally more opaque – hashtags among those related to a given topic as the most innovative and "substandard forms". For each dataset, we evaluated the average z-scores of the 20% most and less common hashtags. The results are displayed in Table 1.

Table 1. Average z-scores of the most common (standard forms) and less common (substandard forms) hashtags

Dataset	Z-scores	
	Standard forms	Substandard forms
BE-1	0.974	-0.145
BE-2	0.450	-0.215
BE-3	1.024	-1.512
BE-4	0.885	0.031
MJ	1.467	-0.024
SF	0.002	0.079

We found that, in all datasets, female users are more likely to use the most common hashtags, which we consider here as the standard forms, than male users. With the exception of the dataset SF, women also use more frequently the standard forms than the substandard ones. This outcome is consistent to the ones reported by the classical sociolinguistic studies.

When analyzing the dataset BE, especially the sub-datasets BE-1 and BE-2 (supporters of the candidates), we could distinguish some hashtags between two categories: those which included personal and direct user's involvement, and those which contained pure voting instructions to the readers of the message. We consider that these two different discursive strategies have the same objective, that is, try to convince readers to vote for a specific candidate. However, linguistically speaking, they are very different, because the usage of a more personal approach brings the reader to a closer position with respect to the author and may be desirable in particular situations.

"Personal involvement" tags are those written in singular 1st person, e.g. #votodilma/#votoserra ("I vote for Dilma/Serra") and #euquerodilma/#euqueroserra ("I want Dilma/Serra"). Persuasive ones are those produced using the 3rd person imperative mood, expressing a command urging the audience to act a certain way, as in #vote13/#vote45 (13/45=Dilma's/Serra's number) and #sejamais1dilma ("Be one more for Dilma"). Table 2 shows the average z-scores of the hashtags of each category.

Table 2. Average z-scores of hashtags including "personal involvement" and "clear persuasion" discursive strategies

Dataset	Z-scores	
	Personal involvement	Voting instruction
BE-1	0.601	-1.894
BE-2	1.477	-0.957

In this case, the "personal involvement" strategy appears to be more common among female users. Male users, on the other hand, tend to be adopters of the "clear persuasion" strategy. Previous studies have found that men are more confident about their ability to persuade, which can be a reason to let them be more confident about using such clear and straightforward strategies. Other studies have also found that female managers, when attempting to convince a subordinate, rely more often on altruism than male managers. Considering that the "personal involvement" strategy reduces the distance between the author and the reader, we could also state that this strategy is related to the altruistic behavior of female managers.

3. CONCLUSIONS AND FUTURE WORK

We present an initial gender based analysis on the usage of hashtags in Twitter. The main purpose of this work is to verify if and how the behavior of male and female users differ in the usage of these specific linguistic elements. This study provides evidence that, although the majority of the hashtags seem to be neutral, some of them are, to some extent, gendered. We have also analyzed different categories of hashtags and found that certain social roles occupied by each gender in offline communities are equally performed in online social networks. Future work will involve the improvement of the statistical analysis and the investigation of other aspects that might act as factors to influence users to employ particular hashtags.

4. REFERENCES

[1] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.P. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. *Proc.* ICWSM, May 2010.

[2] Labov, W. 2001. *Principles of Linguistic Change: Social Factors*. Blackwell, Malden, MA.