Slovene Twitter Analytics

Nikola Ljubešić,[*‡] **Darja Fišer**[†*]
[*] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[‡] Dept. of Information and Communication Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
† Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana, Slovenia
E-mail: nikola.ljubesic@ijs.si, darja.fiser@ff.uni-lj.si

Abstract

The paper presents the results of metadata analysis in a corpus of 7.5 million Slovene tweets. In our analyses we primarily focus on the weekly and daily posting dynamics, their dependence on the account type (corporate vs. private) and user gender, as well as the dependence of the mentioned variables on retweeting, favoriting, text standardness and text sentiment. Through these analyses we gain insight into both user behaviour on social networks and the available linguistic material.

Keywords: Twitter corpus, meta-data analysis, Slovene language

1. Introduction

The large volumes of content generated by Twitter users as well as Twitter's proactive policy have sparked a new venue of research that is attractive for a wide range of disciplines, including information and computer science, media and communication studies, and linguistics. Twitter analytics has been successfully employed to discriminate between different types of users (Mislove et al., 2011) and behaviour (Pennacchiotti and Popescu, 2011; Rao et al., 2010). With state-of-the art techniques, a number of latent user attributes can be identified, such as their location (Hecht et al., 2011), gender (Burger et al., 2011), age (Nguyen et al., 2013), occupation (Hu et al., 2016), social class (Borges et al., 2014) and personality type (Quercia et al., 2011).

This paper is our first attempt at twitter analytics of the Slovene JANES Tweet v0.4 corpus (Fišer et al., 2016a) which contains 7.5 million tweets or 107 million tokens that were posted by nearly 9,000 different users between June 2013 and January 2016. Our goal is to gain insight into user behaviour on social networks and their language characteristics. In addition to the automatically harvested metada during tweet collection, such as posting time, no. of favourites and retweets, the corpus was enhanced with a set of manually and automatically assigned metadata at both user and tweet level. At user level, account type (private / corporate) and user gender (male / female) were manually assigned, while at tweet level text standardness (completely standard / slightly non-standard / very non-standard) and sentiment scores (positive / negative / neutral) were automatically computed.

2. Related Work

Rios and Lin (2013) have used tweet timestamps to visualize annual tweeting dynamics in different cities all over the world, discovering some interesting cultural differences. Scheffler and Kyba (2016), on the other hand, have examined the morning routine of German Twitter users and have found it to be bound to the social norms of working life.

While gender studies on Twitter predominantly focus on gender classification, (Bamman et al., 2012) give a detailed overview of the commonly attributed characteristics of male and female language and behaviour relevant for our study: language standardness (women more standard than men), communication style (men more *informative*, women more *involved*), and characteristic vocabulary (with women exhibiting more distinct features than men, such as frequent use of emoticons, expressive lengthening of words, repeated exclamation marks, etc.).

The typology and granularity of user types varies greatly in the literature. While they typically exceed the two classes used in our corpus, most researchers distinguish *organizations*, such as news media outlets and public institutions from other users. Arakawa et al. (2014) have the closest reading to our *corporate users* in their *organizations* category, which they were able to classify with the highest accuracy. They report that tweets from organizations posted the highest number of tweets the objective of which is to transmit information, which is characterized by a distinctly high use of nouns, polite language, hashtags, URLs and retweets.

The relationship between gender and subjective language in tweets has been explored for English, Spanish and Russian by Volkova et al. (2013) who have shown that there are substantial differences in the use of subjective words (e.g. *weakness*, which is used to express positive sentiment by women and negative by men), hashtags (e.g. *baseball*, which expresses positive sentiment by men and negative by women) and emoticons (with women using more emoticons overall than men in English and Spanish but, interestingly, not in Russian) and that these differences can improve sentiment classification.

3. Posting Dynamics

The first part of our statistical analyses focuses on the volume of posts, retweets and favourites. We inspect the

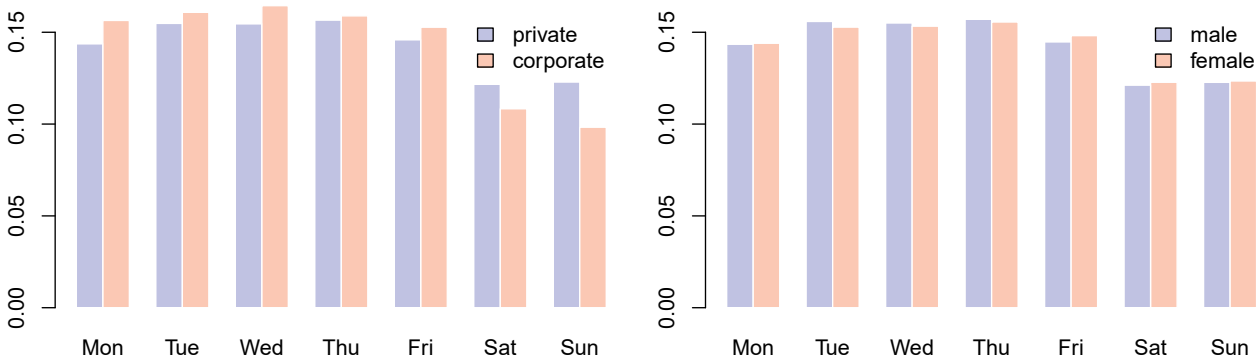

Figure 1: Probability distribution of tweets by day of week, separate by source (left) and gender (right).

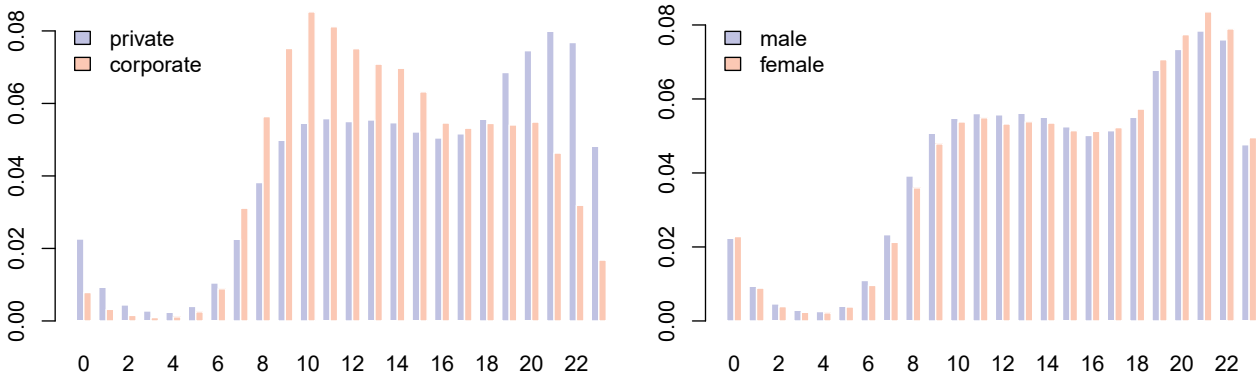

Figure 2: Probability distribution by hour in day, separate by source (left) and gender (right).

weekly and daily posting cycles and their dependence on the account type (private vs. corporate) and user gender (male vs. female). Finally, we inspect the dependence of the two last variables and post retweeting and favouriting.

3.1. Weekly Posting Cycle

The weekly posting cycle is presented in Figure 1 where the graph on the left shows distributions for private and corporate accounts while distributions for male and female users of private accounts are displayed on the right. We can see that while the overall volume of tweets posted is higher on weekdays, corporate users are dominant during the week and private ones on weekends which is not surprising but does have important implications on the topics and the language of the tweets published during the week vs. on weekends. Genderwise the distributions are very similar to the type of user, with male users prevailing mid-week and females on weekends.

3.2. Daily Posting Cycle

Figure 2 shows the daily posting cycle with user behaviour per account type displayed on the left and behaviour per user gender limited to private accounts on the right. As expected, tweeting volume of corporate users peaks during morning hours (11 a.m.) while private users are most active in the evening (9 p.m.). Interestingly, both types of users have a secondary peak that coincides with the period of the major peak of the other group. In terms of user gender, male users dominate slightly from 1 a.m. to 3 p.m. after which female users take over and are more active throughout the afternoon and evening, suggesting that male users display behaviour a bit similar to corporate accounts while females display a distinct private-use behaviour tweeting in their spare time after work.

	retweeted	favorited
private	8.5%	30.2%
corporate	16.3%	18.0%
male	9.4%	29.2%
female	6.8%	32.9%

Table 1: Probabilities of tweets to be retweeted, i.e. favorited, given account type and user gender variables.

3.3. Retweets and Favorites

Next we make comparisons between the retweeted and favorited variables on one side and the source and gender variables on the other. We operationalise the retweet and favorite variables as binary variables that are true if a tweet was retweeted or favorited, respectively. We present the percentages of the retweeted or favourited tweets given the account type or user gender in Table 1.

We begin by inspecting the dependence of the source variable and the retweet variable. The probability of a corporate tweet to be retweeted is twice as high as for private tweets, which was to be expected as the primary function of most corporate tweets is information dissemination. Running the chi-square test of independence proves for the vari-

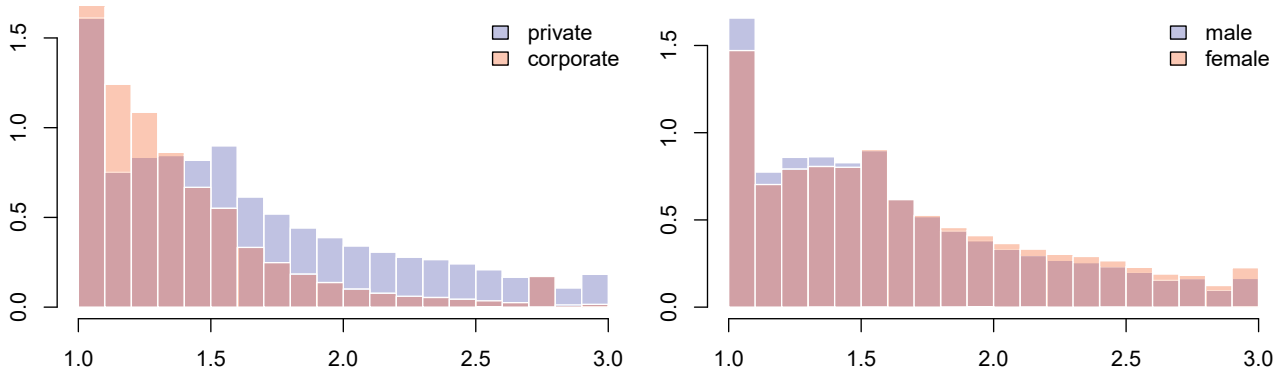

Figure 3: Distribution of the three standardness levels by account type (left) and user gender (right). Lilac represents distribution overlap.

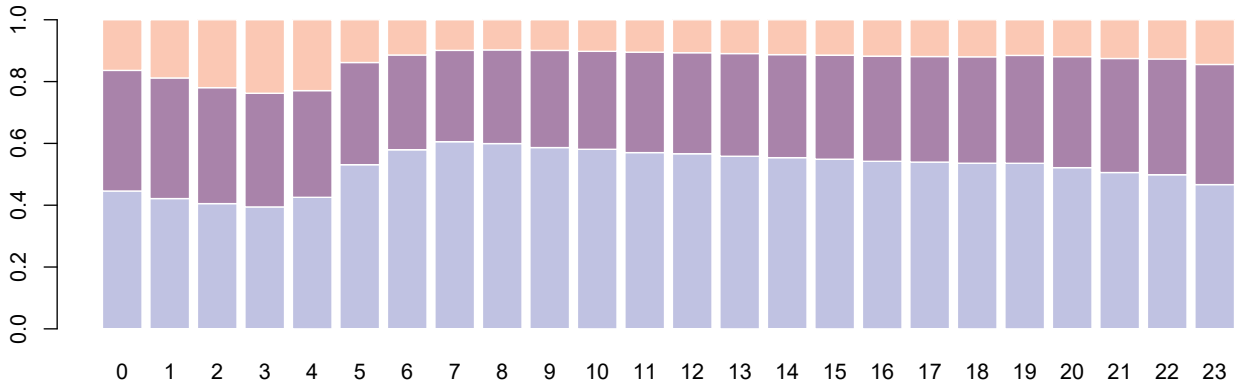

Figure 4: Standardness by hour of day, standard represented with blue, slightly non-standard with lilac, very non-standard with red.

ables of source and retweets not to be independent with $X^2(1, N = 7503200) = 74308, p < .001$.

Similarly, analysing the dependence of the source variable and the favorite variable, we measure that private tweets tend to be almost twice as frequently favorited as corporate tweets which is again consistent with the communicative role of private posts that have a strong community- and relationship-building role. The chi-square test of independence shows a relationship between the source and favorite variable with $X^2(1, N = 7503200) = 80215, p < .001$.

Moving to the comparison with the gender variable, we first inspect the dependence of the gender and the retweets variable. Male tweets are 38% more probable to be retweeted than female tweets. Calculating the chi-square test of independence shows a relationship between these two variables with $X^2(1, N = 7503200) = 11714, p < .001$.

By comparing the gender and favorited variables, we calculate that it is 13% more likely for a female tweet to be favorited than a male tweet. The chi-square test of independence shows a relationship between the gender and favorite variable with $X^2(1, N = 7503200) = 8913.4, p < .001$.

The presented results again suggest that male Twitter users behave more like corporate users and females are more aligned with the private Twitter accounts.

4. Language Standardness

The second part of statistical analyses inspects the linguistic characteristics of tweets posted by the different groups of users. Due to space constraints, we only present the results for language standardness scores assigned to each tweet in the corpus via a regression model (Ljubešić et al., 2015) while the behaviour of tweets according to the percentage of normalised tokens via CSMT (Ljubešić et al., 2016) that was also computed is consistent with the text standardness results.

We inspect the relationship of the account type and the user gender variable on one hand and the standardness continuous variable (ranging from 1 to 3) on the other. The resulting plot is presented in Figure 3. We can see that tweets posted by private and corporate users differ significantly regarding linguistic standardness, corporate users showing a much stronger tendency towards standard language, which is not surprising given their communicative goal. Male and female users are much more similar in this respect, but male users tend to produce more standard tweets, while female ones produce more semi- and non-standard ones, which is an interesting finding that deserves a closer examination in future work.

Given that the difference in text standardness by user gender presented in the right plot of Figure 3 is minor, we perform the chi-square test of independence showing a relationship of user gender and tweet standardness with

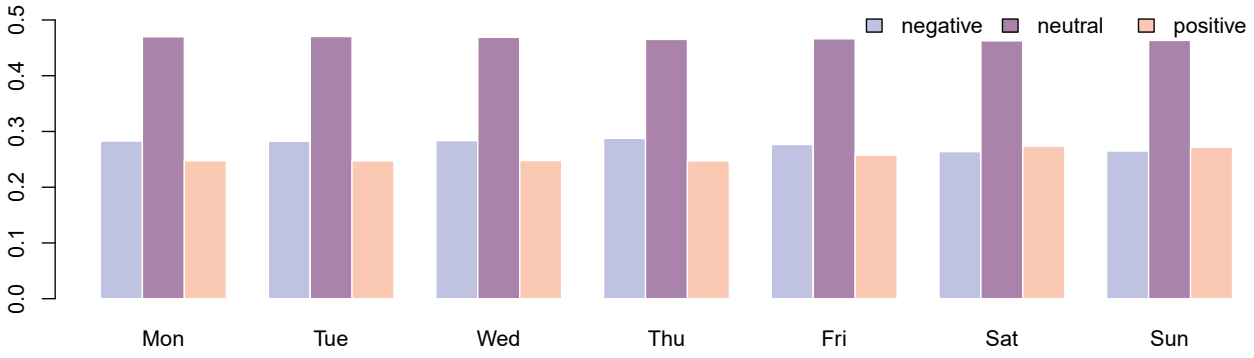

Figure 5: Distribution of the sentiment by day of week among private users.

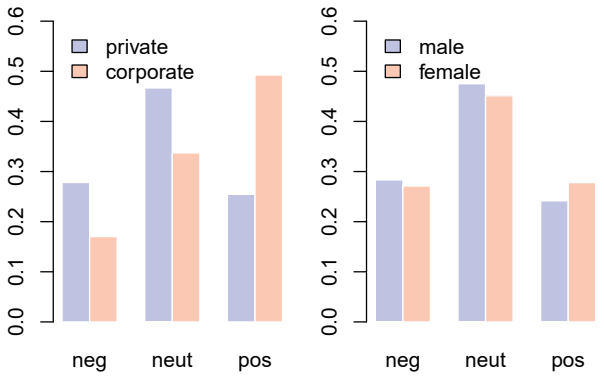

Figure 6: Distribution of the sentiment by source (left) and gender (right).

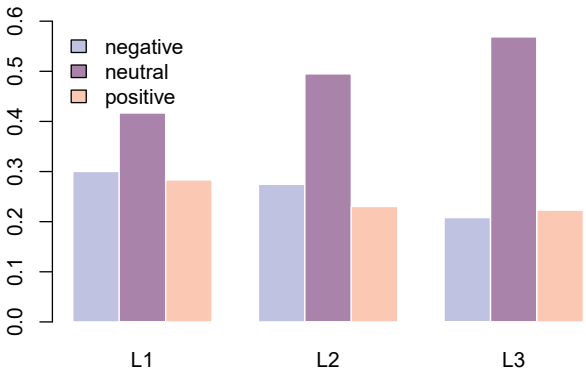

Figure 7: Distribution of the sentiment by language standardness (L1 - completely standard, L2 - slightly non-standard, L3 - very non-standard).

$X^2(1, N = 7503200) = 9740.9, p < .001$. For this test we operationalise the standardness variable as a binary variable, discarding tweets that are by the discrete standardness variable (three levels) slightly non-standard. While 24% of the remaining female tweets are estimated as being non-standard, for males the percentage is 19.6%.

Next, we plot the distribution of the discrete standardness variable (three levels) in the daily posting cycle in Figure 4. As expected, tweets are the most standard in the early morning hours (7 a.m.), which is probably an effect of corporate accounts of newspapers and other media posting links to new content for the day. As the day progresses, the proportion of slightly non-standard tweets rises steadily as does the proportion of very non-standard ones but they go up only slightly until late evening hours (after 11 p.m.) when they pick up and peak at around 3 a.m.

5. Sentiment Analysis

Finally, we look into the relationship of the account type and user gender variables with the sentiment score automatically assigned to each text in the Janes corpus using SVM (Fišer et al., 2016b). The three variables are compared in Figure 6. While corporate users post predominantly positive tweets and private users more neutral and negative ones, male users post slightly more negative posts and female users take the lead in the positive ones.

Again, given the close results on user gender, we perform the chi-square independence test of the user gender variable and the binary positive / negative text sentiment variable, discarding thereby neutral tweets. The test shows a relationship between the gender and the sentiment variables with $X^2(1, N = 7503200) = 6179.8, p < .001$.

In order to gain more insight into how the sentiment of Slovene tweet users varies throughout the week, we plotted the relationship of posting day and tweet sentiment in Figure 5. Disregarding the neutral tweets which prevail every day of the week, we can see that users start the week with a distinctly negative attitude which peaks on Thursday and then starts decreasing on Friday so that positive sentiment prevails during the weekend, peaking on Saturday.

The relationship between sentiment and standardness among private users is examined in Figure 7. Disregarding the neutral tweets that are prevalent across the board, positive sentiment prevails in very non-standard posts while the opposite is true at the other end of the spectrum. Our plan is to investigate this dependence in more detail in future work.

6. Conclusions

In this paper we carried out an analysis of a series of extralinguistic and linguistic variables in a large corpus of Slovene tweets. Among many of our findings, the most interesting ones are that there are big differences between tweeting behaviour, content and treatment of corporate and private tweets that are aligned with the primary communicative functions of the two types of Twitter users. Pri-

vate male users tweet more than female users during weekdays while female users dominate on weekends. Male users tweet more in the morning hours while female users take the lead in the afternoon and evening. Male users use more standard language than female users, which is most frequently used in the early morning hours overall. Female users express more positive sentiment in their posts than their male counterparts, which is the prevalent sentiment overall while both tend to be more positive on weekends than during the week.

While the results are difficult to compare directly with the related work, the results obtained for the communication behaviour and styles of private and corporate users closely resemble the ones reported by Arakawa et al. (2014), Scheffler and Kyba (2016) and Volkova et al. (2013). The most striking difference between our results and related work is the language standardness level, which is higher in male users, contrary to what Bamman et al. (2012) have observed.

In the future we plan to extend our work with comprehensive statistical content and linguistic analyses. We also wish to compare the results with other text genres in the JANES corpus, such as blog posts, forum messages, news comments and Wikipedia talk pages. Finally, we envisage to compare the results with similar languages, such as Croatian and Serbian.

7. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Nonstandard Internet Slovene" (J6-6842, 2014-2017).

8. References

Arakawa, Y., Kameda, A., Aizawa, A., and Suzuki, T. (2014). Adding twitter-specific features to stylistic features for classifying tweets by user type and number of retweets. *Journal of the Association for Information Science and Technology*, 65(7):1416–1423.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2012). Gender in Twitter: Styles, stances, and social networks. *CoRR*, abs/1210.4567.

Borges, G. R., Almeida, J. M., Pappa, G. L., et al. (2014). Inferring user social class in online social networks. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, page 10. ACM.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fišer, D., Erjavec, T., and Ljubešić, N. (2016a). The compilation, processing and analysis of the JANES corpus of Slovene user-generated content. *Slovenščina 2.0*, 4(2):67–100.

Fišer, D., Smailović, J., Erjavec, T., Mozetič, I., and Grčar, M. (2016b). Sentiment Annotation of the Janes Corpus of Slovene User-Generated Content. In *Proceedings of the 10th conference on language technologies and digital humanities*.

Hecht, B., Hong, L., Suh, B., and Chi, E. H. (2011). Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA. ACM.

Hu, T., Xiao, H., Luo, J., and vy Thi Nguyen, T. (2016). What the Language You Tweet Says About Your Occupation.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S., and Škrjanec, I. (2015). Predicting the Level of Text Standardness in User-generated Content. In *Proceedings of Recent Advances in Natural Language Processing*, pages 371–378.

Ljubešić, N., Zupan, K., Fišer, D., and Erjavec, T. (2016). Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of KONVENS 2016*.

Mislove, A., Jørgensen, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J., (2011). *Understanding the Demographics of Twitter Users*, pages 554–557. AAAI Press.

Nguyen, D.-P., Gravel, R., Trieschnigg, R., and Meder, T. (2013). " how old do you think i am?" a study of language and age in twitter.

Pennacchiotti, M. and Popescu, A.-M. (2011). A machine learning approach to twitter user classification.

Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 180–185. IEEE.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA. ACM.

Rios, M. and Lin, J. (2013). Visualizing the "pulse" of world cities on twitter.

Scheffler, T. and Kyba, C. C. (2016). Measuring social jetlag in twitter data. In *Tenth International AAAI Conference on Web and Social Media*.

Volkova, S., Wilson, T., and Yarowsky, D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1815–1827.