Socialbots supporting human rights

E. Velázquez, M. Yazdani, P. Suárez-Serrato
IMUNAM, CALIT2 UCSD, IMUNAM

Abstract

Socialbots, or non-human/algorithmic social media users, have recently been documented as competing for information dissemination and disruption on online social networks. Here we investigate the influence of socialbots in Mexican Twitter in regards to the "Tanhuato" human rights abuse report. We analyze the applicability of the BotOrNot API to generalize from English to Spanish tweets and propose adaptations for Spanish-speaking bot detection. We then use text and sentiment analysis to compare the differences between bot and human tweets. Our analysis shows that bots actually aided in information proliferation among human users. This suggests that taxonomies classifying bots should include non-adversarial roles as well. Our study contributes to the understanding of different behaviors and intentions of automated accounts observed in empirical online social network data. Since this type of analysis is seldom performed in languages different from English, the proposed techniques we employ here are also useful for other non-English corpora.

As of 2017, Twitter has over 318 million monthly "active users" (Sparks 2017) - an amount that is more than population of Indonesia, the 4th most populous country in the world. Advances in Artificial Intelligence, however, has made it possible to automate the creation of online social media accounts that attempt to behave similarly to human users. These non-human accounts are known as *socialbots* or simply bots. A range of intentions and goals drive the production and deployment of such bots. Often, socialbots intervene in the discussion of specific trending topics to potentially manipulate, deceive, and distract human users (for one review, see (Ferrara et al. 2016)).

While there have been numerous studies on the impact and influence of socialbots, most previous studies have been limited to English Twitter. In this paper we present a case study to see how social bots are used in Mexican Twitter on a specific trending topic. We followed over 20 social events in Mexican Twitter in 2016, covering topics ranging from political scandals, attacks against the media, journalists, expressions of homophobia, to the banal and trivial. We found that the topic related to the report documenting the violation of human rights in Tanhuato had far more bot activity than other topics. We therefore focus our scope of study to tweets related to the #Tanhuato hashtag that was trending in relation to the release of this report. We now give a brief background of this hashtag and topic.

Background

As part of the war on drugs on May 22nd, 2015 the Mexican armed forces raided a ranch in Tanhuato, Michoacán. After an extensive investigation, the National Commission for Human Rights (Comisión Nacional de Derechos Humanos CNDH) released a report on August 18th, 2016. They established that at least 22 civilians were arbitrarily executed, victims suffered instances of torture, and that the crime scene was tampered with.

The report from the CNDH was made available online, and they used their Twitter account to promote access to it [1]. During the following days there was an increased interest in the topic in Mexican Twitter, using the hashtag #Tanhuato. We collected over 20K tweets using Twitter's streaming API between the 19th and 21st of August 2016. These tweets were processed and the user ID's evaluated with *BotOrNot* (Davis et al. 2016) immediately after collection[2].

We found a substantive presence of socialbots using the #Tanhuato hashtag during the collection period. According to *BotOrNot*(Davis et al. 2016), out of a total of 9,730 unique accounts we found high bot scores for 1,777 accounts. By following the retweets of the total collection of users we found an additional 26 bot accounts, giving us a total of 1,803 bots detected. Given this significant bot activity, we investigate what is the intention behind such bot accounts and their impact on spreading or stifling information. Since most text and bot analysis is typically done with English corpora, we also adapt our analysis for tweets in Spanish.

Unexpectedly, we found from our analysis that in fact most of the bot accounts were not acting maliciously and were in fact promoting access to the CNDH report. Human users retweeted the content of the tweets generated by bots, so that access to this report proliferated through the support of the socialbots and in coordination with the human users that retweeted them. What was the intention of the bots that

[1] Report available at
http://www.cndh.org.mx/sites/all/doc/
/Recomendaciones/ViolacionesGraves/RecVG_004.pdf

[2] Predecessor of *Botometer*.

we detected using #Tanhuato ? We shall argue that they were helping to provide access to the report issued by the CNDH. This type of behavior sets them apart from the typically observed bots that have spam, or even censorship, intentions (Woolley 2016), (King, Pan, and Roberts 2013).

It is important to pause here and notice that in an instance like this it is not a clear matter whether these bots were benevolent or malignant. It is a matter of perspective. From the point of view of the Mexican armed forces, these bots are acting against their honor. From the point of view of the CNDH they are promoting access to a report of human rights abuse. Our study thus provides an interesting empirical test case for social bots acting as promoters, as opposed to suppressors, of information.

Previous work

Correlations of content between different accounts has also been used as a twitter bot detection technique (Chavoshi, Hamooni, and Mueen 2016). Tweet sentiment has been studied to discriminate human from non-human accounts (Dickerson, Kagan, and Subrahmanian 2014). Other methods combine graph-theoretic, syntactic, and semantic features to find bots (Chu et al. 2010). Another method to identify bots exploits natural language processing (Clark et al. 2016). The possibility of creating a call to arms for activists using Twitter has been previously explored, and in fact this case study seems to be a variation on this theme (Savage, Monroy-Hernandez, and Höllerer 2016). Numerous other previous works have addressed the issues of detection and classification of bots, see for example (Wang 2010; Dickerson, Kagan, and Subrahmanian 2014; Hu et al. 2013; Thomas et al. 2011; Yang et al. 2011; Zhu et al. 2012; Lee, Eoff, and Caverlee 2011; Ratkiewicz et al. 2011; Thomas et al. 2013; Lee and Kim 2014; Beutel et al. 2013; Hu et al. 2014; Boshmaf et al. 2013).

Most of the mentioned methods and previous results have been developed for English. By using the language-independent features of *BotOrNot* (Davis et al. 2016) it is possible to flag potential bot accounts in Spanish, and in other languages as well.

Bot identification and data preparation

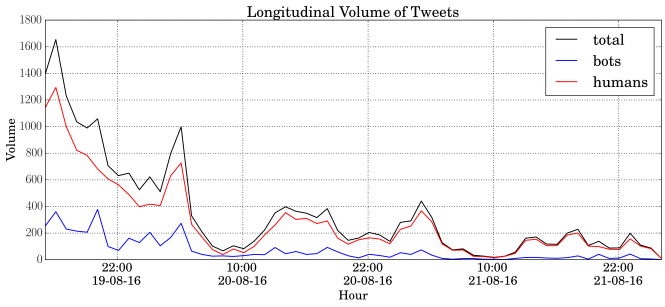


Figure 1: Bot versus human activity using #Tanhuato, from streamed tweets during collection period.

In order to detect bots we use *BotOrNot*, a general supervised learning system designed for detecting socialbot accounts on Twitter (Davis et al. 2016). It utilizes over 1,000 features such as user meta-data, social contacts, diffusion networks, content, sentiment, and temporal signatures. Based on evaluation on a large set of labeled accounts, *BotOrNot* is extremely accurate in distinguishing bots from humans accounts, with an Area Under the ROC Curve (AUC) of 94%.

When a twitter account is evaluated in *BotOrNot*, the output is a JSON file with several scores. As we are examining a corpus of tweets in Spanish we focus on language-independent classifiers, which show a large number of potential bot accounts. Surprisingly, combining the results of these language-independent classifiers is sufficient for detecting bots in Spanish. This suggests that simply discarding the language-dependent features of *BotOrNot* can yield to non-English bot detection. Further research should be done to validate the transferability of *BotOrNot* outside of English Twitter.

We streamed 20,854 tweets from Twitter's API between 2016-08-19 15:06:17 and 2016-08-22 02:13:35. These tweets were generated by 9730 different users (see Figures 1 and 5 for the relation between humans and bots), and among them we have 12905 retweets. When a user (human or bot) generates a tweet, and this tweet can be retweeted by a bot or a human. Consequently, we find four possibilities: a tweet created by a human and retweeted by another human (**H-H**), created by a human and retweeted by a bot (**H-B**), created by bot and retweeted by human (**B-H**) or bot (**B-B**). In Figure 1 we show the evolution of #Tanhuato in the collection period. The percentages of accounts that are humans and those that are bots are shown in Figure 4.

In Figure 2 we show the bi-variate kernel decomposition estimates for pairwise combinations of the Friend, Network, and Temporal classifiers form *BotOrNot*. The regions towards the upper right hand corner correspond to areas where the bot scores are high. It can be clearly seen how the bot accounts naturally cluster. The final visualization of this analysis is presented in Figure 3, where we now compute the kernel density estimate that incorporates the three classifiers Friend, Network, and Temporal. In this image the smaller cluster in the upper right corner is the region where the bots accumulate. This 3D image is formed by taking iso-surfaces obtained from the 3D kernel density estimate. Again, as in the 2D images, we can separate the bot accounts in a natural way, to isolate them for further analysis. Notice that these three classifiers are all non-language specific and this is the reason behind focusing on them instead of on the overall bot score produced by *BotOrNot*. Having identified the bots present in our sample, we can now understand how the appeared over the collection period, as shown in Figure 4.

Network Analysis

Now that we have performed our bot analysis, we can analyze the bot and human Twitter network. In Figure 6a we see that the nodes with the highest betweenness centrality in the full retweet network are all human, except for two accounts that belong to bots. These bot accounts are in fact official news organizations @pictoline and @Pajaropolitico. Thus,

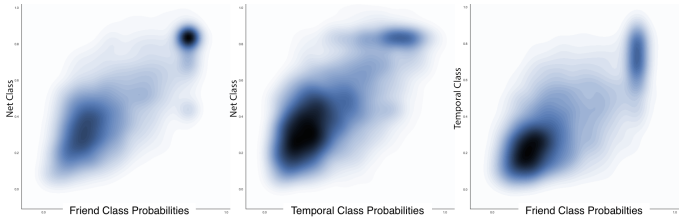


Figure 2: 2D Kernel decomposition estimate for Friend, Network, and Temporal classifiers from Bot-Or-Not, for #Tanhuato, 19-21st August 2016, sample obtained through Twitter's streaming API.

by the betweenness centrality in the retweet network, human users constitute the shortest paths of dialogue. With the exception of the formal news bots, socialbots are not playing an active role in the retweet network.

Figure 6b shows the number of retweets by each user (measure of degree in the retweet network) and that again humans are the more active retweeters. In Figure 5 (DOWN) we find the relation between these quantities for our data. Furthormore, we observed in the data that the bots with the highest ammount of retweets among humans were mainly news organizations: @pictoline, @Pajaropolitico, @emeequis, @CNNEE, and @NewsweekEspanol.

Text Analysis

We extract bag-of-words features represented as TF-IDF (term frequencyinverse document frequency) using (Buitinck et al. 2013). We then used Singular Value Decomposition (SVD, also referred to as Latent Semantic Indexing in the context of information retrieval and text mining) to look at the distribution of Tweets on the top singular vectors. While the top singular vectors capture the most variance in the bag-of-words features set, for this corpora the difference between the bot and human tweets was not clear. We also redid the analysis by removing Spanish stop words and still did not find any discrimination between bots and humans.

To better understand the nature of words bots and humans used, we apply basic sentiment analysis using LabMT (Dodds et al. 2011). As discussed in (Dodds et al. 2011), the top 10,000 Spanish words were presented to Amazon Mechanical Turk where 50 workers rated the happiness of each word on a scale of 1 to 9 (where 1 is least happy, 9 is most happy, and 5 is neutral). Using these scores for each word, we compute the average sentiment, h_{avg} for the human and bot corpora using Equation 1 in (Dodds et al. 2011). As discussed in (Dodds et al. 2011) however, a great deal of words may have neutral sentiment (and are essentially commonly used stop words), and the average sentiment score may be biased heavily towards the neutral score of 5.0. Therefore, the authors suggest removing words that are within Δh_{avg} of 5.0 so that words with stronger sentiment remain. By selecting an appropriate Δh_{avg}, we can remove stop words in a systematic way that does not contribute to sentiment.

It is not clear what value to select for Δh_{avg}. While the authors in (Dodds et al. 2011) suggest $0.5 \leq \Delta h_{avg} \leq 2.5$, here we compute the average sentiment score Δh_{avg} for $0 \leq \Delta h_{avg} \leq 3.0$ for a more complete understanding. Figure 7, left panel, shows how the tweets average sentiment changes as we filter out more neutral words. As the neutral words are filtered, we see that the average sentiment is pulled down significantly. This is to be expected as most tweets are expressing words related to violence. Interestingly, however, the bots seem to be less emotional than the humans in that their average sentiment is consistently above humans regardless of what Δh_{avg} value we use.

To investigate this hypothesis further, we removed all retweets and recomputed the average sentiments. Figure 7, panel on the right, shows again that removing the retweets does not change the fact that filtering neutral words yields more negative words. However, we see that the bot sentiment does not correlate strongly with the human tweets. In other words, as we filter more neutral words, the human tweets become more negative as before. But the bot tweets remain closer to being neutral. These findings all suggest that the bots were using less emotionally charged words than humans. In other words, it appears that the purpose of the bots in this case was to only distribute information in a non-sensational manner rather than purposefully stir up emotions.

In addition to using LabMT, we also hand coded a list of negative words, extracted from the corpus of collected tweets, and used it to compare both the bot and human corpora according to the frequency of appearance of words in this list. In order to increase the comparability of these words in a wider volume of tweets, when possible, we suppressed some last letters (that is, we applied "stemming") such that they could match with different tenses (in case of verbs) and different genders and numbers (in nouns and adjectives) keeping the connotation. We refer to Table 1 for this list of incomplete words.

To check matches between words in Table 1 and the text in tweets, we remove URLs from the text in tweets, replace non-ASCII characters (like "ñ ", stressed vowel á,é,í,ó,ú and "¿ ") by their ASCII equivalent ("n",a, e, i, o, u,"?"). We also transform all capital letters to lowercase. The transformed text were split into single words to compare individually. In order to increase comparison speeds, we group the words alphabetically and compare only with words starting with the same letter, skipping also words starting with symbols, numbers. Finally, we only check if the words in Table 1 with the same initial letter as each word in split message starts with the same letters.

To prevent a misplaced punctuation mark from not matching a word, a second analysis was performed suppressing the first letter in each word, and checking if this shorter word matches with Table 1. This analysis also reveals no difference. Our method of comparison fails when a negative sentiment word is misspelled, but one expects that the sentiment of the tweet remains congruent in the whole text. Then, if the text is long, we are more likely to find another negative word but spelled correctly. Conversely, short texts are more likely to have less misspelled words.

To distinguish what kind of information is most shared, we consider the total of tweets and assign a numerical value

arma	culpable	jodid∗	sanguinari∗
asesin∗	delincuen∗	levanton	secuestro
asesinat∗	dispara	maltrat∗	tortura
bala	disparos	masacre	violacion
balazo	ejecucion	matanza	violenta
brutal	ejecut∗	matar	
cartel	exterminio	mentir	
castigo	fals∗	muerte	
corrupcion	genocidio	pistola	
corrupt	guerra	represion	
crimen	incendia	represiv∗	
criminal	jode∗	sangriento	

Table 1: List of negative feeling words (an ∗ is placed when letters can omitted without changes in connotation.

to each one. This value was initialized in 0 increased by a constant, depending on the number of matches with the Table 1. Assuming that a tweet has a negative feeling when its value is different to zero, we show in Figure 8 that the largest volume of tweets comes from retweets with a negative feeling text. A closer reading of the entire tweet corpus revealed that the most of the messages which are non-negative cannot be identified as positive or neutral. Their texts share URLs and/or the sentiment cannot be determined by word inspection.

Conclusions

In this work we presented a case study of socialbots for a specific trending topic in Mexican Twitter. While numerous studies have suggested that socialbots act as disrupting agents of information, in our case study we found the opposite. The socialbots were in fact enabling the flow of information to ensure that the report about these atrocities reached the public and information was not stifled. Of course, from the point of the police authorities the socialbots may be viewed as agents of disruption and it is therefore a matter of perspective if socialbot are enablers or not. Our case study suggests that the role and landscape of socialbots is far more complex than simple binary categorizations. Our work highlights the need for further research to understand the ethical implications of such automated social activity.

Acknowledgments.

We thank IPAM in UCLA and the organizers of the Cultural Analytics program, CNetS and the BotOrNot team in IU, and also Twitter for allowing access to data through their APIs. PSS acknowledges support from UNAM-DGAPA-PAPIIT-IN102716 and UC-MEXUS CN-16-43.

References

Beutel, A.; Xu, W.; Guruswami, V.; Palow, C.; and Faloutsos, C. 2013. Copycatch: Stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, 119–130. New York, NY, USA: ACM.

Boshmaf, Y.; Muslukhov, I.; Beznosov, K.; and Ripeanu, M. 2013. Design and analysis of a social botnet. *Comput. Netw.* 57(2):556–578.

Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; Layton, R.; VanderPlas, J.; Joly, A.; Holt, B.; and Varoquaux, G. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122.

Chavoshi, N.; Hamooni, H.; and Mueen, A. 2016. Identifying correlated bots in twitter. In *International Conference on Social Informatics*, 14–21. Springer International Publishing.

Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2010. Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, 21–30. New York, NY, USA: ACM.

Clark, E. M.; Williams, J. R.; Jones, C. A.; Galbraith, R. A.; Danforth, C. M.; and Dodds, P. S. 2016. Sifting robotic from organic text: A natural language approach for detecting automation on twitter. *Journal of Computational Science* 16:1 – 7.

Davis, C. A.; Varol, O.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, 273–274. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Dickerson, J. P.; Kagan, V.; and Subrahmanian, V. 2014. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* 00(undefined):620–627.

Dodds, P. S.; Harris, K. D.; Kloumann, I. M.; Bliss, C. A.; and Danforth, C. M. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one* 6(12):e26752.

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

Hu, X.; Tang, J.; Zhang, Y.; and Liu, H. 2013. Social spammer detection in microblogging. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, 2633–2639. AAAI Press.

Hu, X.; Tang, J.; Gao, H.; and Liu, H. 2014. Social spammer detection with sentiment information. In *Proceedings of the 2014 IEEE International Conference on Data Mining*, ICDM '14, 180–189. Washington, DC, USA: IEEE Computer Society.

King, G.; Pan, J.; and Roberts, M. E. 2013. How censorship in china allows government criticism but silences collective expression. *American Political Science Review* 107(2 (May)):1–18. Please see our followup article published in

Science, “Reverse-Engineering Censorship In China: Randomized Experimentation And Participant Observation.”.

Lee, S., and Kim, J. 2014. Early filtering of ephemeral malicious accounts on twitter. *Comput. Commun.* 54(C):48–57.

Lee, K.; Eoff, B. D.; and Caverlee, J. 2011. Seven months with the devils: a long-term study of content polluters on twitter. In *In AAAI Intl Conference on Weblogs and Social Media (ICWSM*.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonalves, B.; Patil, S.; and Flammini, R. 2011. Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th interna)onal conference companion on World wide web*.

Savage, S.; Monroy-Hernandez, A.; and Höllerer, T. 2016. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW ’16, 813–822. New York, NY, USA: ACM.

Sparks, D. 2017. How many users does twitter have? [Online; accessed 06-June-2017].

Thomas, K.; Grier, C.; Song, D.; and Paxson, V. 2011. Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC ’11, 243–258. New York, NY, USA: ACM.

Thomas, K.; McCoy, D.; Grier, C.; Kolcz, A.; and Paxson, V. 2013. Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. In *Proceedings of the 22Nd USENIX Conference on Security*, SEC’13, 195–210. Berkeley, CA, USA: USENIX Association.

Wang, A. H. 2010. *Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach*. Berlin, Heidelberg: Springer Berlin Heidelberg. 335–342.

Woolley, S. 2016. Automating power: Social bot interference in global politics. *First Monday* 21(4).

Yang, Z.; Wilson, C.; Wang, X.; Gao, T.; Zhao, B. Y.; and Dai, Y. 2011. Uncovering social network sybils in the wild. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC ’11, 259–268. New York, NY, USA: ACM.

Zhu, Y.; Wang, X.; Zhong, E.; Liu, N. N.; Li, H.; and Yang, Q. 2012. Discovering spammers in social networks. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI’12, 171–177. AAAI Press.

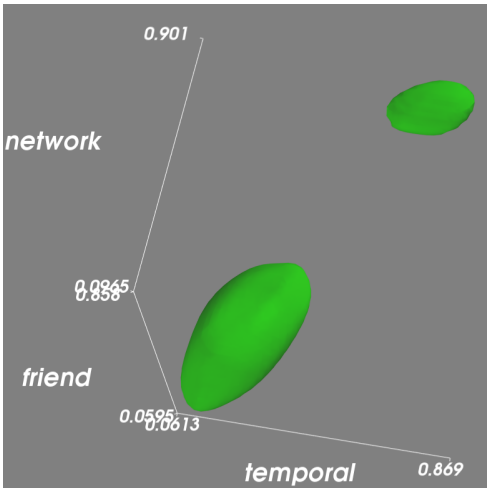


Figure 3: 3D Kernel decomposition estimate for Friend, Network, and Temporal classifiers from Bot-Or-Not, for #Tanhuato, 19-21st August 2016, sample obtained through Twitter’s streaming API.

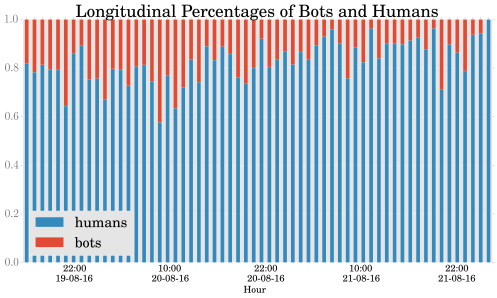


Figure 4: Percentages of Bot versus human activity using #Tanhuato, from streamed tweets during the collection period.

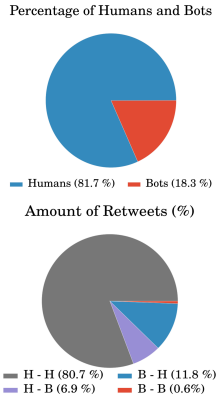
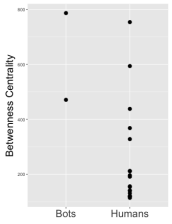
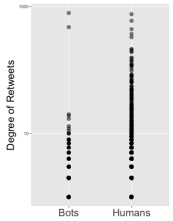


Figure 5: UP: Percentage of different human and bot accounts in collected data. Volume of registered retweets by user type. DOWN: Clasification is as follows: humans retweeting humans (**H-H**), bots retweeting humans.



(a) Retweet network betweenness centrality: the two bots are news organizations.



(b) Highest degree nodes (raw retweet counts) in the retweet network.

Figure 6: Distribution of centrality of bot and human Twitter accounts. . We only show the top Twitter accounts.

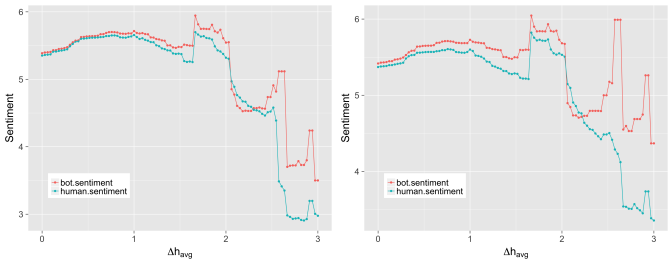


Figure 7: **Left:** Sentiment on Tweets using LabMT. As we filter out neutral words with the Δh_{avg}, we see that the sentiment from human is significantly lower than bots.. **Right:** Sentiment on Tweets with retweets removed using LabMT. Again, as we filter out neutral words with the Δh_{avg}, we see that the sentiment from human is significantly lower than bots. However, the correlation between the human and sentiments is much lower when retweets are removed.

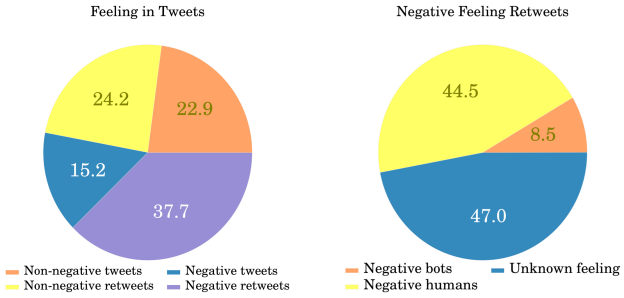


Figure 8: The total volume of twitter texts were comparing with words in Table 1. LEFT: Tweet classification in **Negative** and **Non-negative**. RIGHT Percentage of negative feeling texts by user type.