December 1, 2015

Abstract

1 Introduction

One of the most interesting border regions in the world is composed by the cities of San Diego, California, USA and Tijuana, Baja California, Mexico. As of 2014, it had a combined population of over 40 million people, making it the biggest metropolitan border region on the US-Mexico frontier. The physical proximity of these cities has led sociological researchers to propose them as a new urban formation: the trans-border metropolis. This formulation proposes a functional metropolitan area that physically transcend international borders, and where urban management in such areas can only be achieved through a combination of city planning and international diplomacy. On the physical level, multiple roads and highways closely link these cities. Thus, from the air, this region can be considered as one continuous urban agglomeration. On a behavioral level, there is a significant amount of work, shopping, social and touristic integration of the populations. However, there is minimal integration on the communications level, only limited cultural exchanges, and there is little to no bi-national integration on a politico-administrative level.

2 Data and Methodology

We collected 10,908,817 tweets using Twitter's Streaming API[1] starting from December 4th, 2013 until January 13, 2015. A bounding box (32 25' 4.2414" N, 117 18' 49.5066" W and 33 5' 53.3178" N, 116 49' 17.9142" W) was used to filter only to those messages originating from San Diego and Tijuana. Each tweet consisted of a unique identifier, date-time of submission, coordinates

[1] https://dev.twitter.com/

of submission (i.e., latitude and longitude), language of submission, and the message itself. Even though this process comes with the limitation that we only have access to approximately 1% of all the tweets [6], previous work have shown that the data obtained from the API closely resembles a random sample drawn from the full Twitter stream [5].

Each tweet was preprocessed in the following way: (i) all characters were lowercased; (ii) URLs and mentions were replaced with placeholders __URL__ and __MENTION__ receptively; (iii) Four-digit numbers were replaced by __4NUM__, any other length number was replaced by __NUM__, (iv) Every punctuation character was removed except for @,#,_ and new-lines; (v) negations were handled by marking all tokens up to the next punctuation. Tweets were grouped according to their submission language. This resulted in two datasets, one for English ($12,212,416$ tweets) and one in Spanish ($764,709$ tweets).

2.1 Obtaining country of submission

In order to find the country of submission from the GPS coordinates, we trained a Gradient Boosting Classifier (GBM). GBM is a random forest ensemble method based on the decisions of weak tree classifiers. This method has been shown to produce good results for most problems [2]. We trained our classifier on 2838 coordinates-country pairs obtained from the GeoNames Gazetteer[2]. Our model was trained to classify latitude and longitude into country names with options between U.S., M.X., and Other. We tested the model's accuracy in a held-out test set of 500 samples obtaining 99.18% accuracy.

2.2 Extracting sentiment

The sentiment analysis method used was a simplistic variation of the SAIL method, presented in [3] for the SEMEVAL 2014 Challenge. In its original implementation, it is a method that derives lexicon features that describe the emotion of a text through a multitude of statistics.

2.2.1 Expanding dictionaries

To obtain comparable metrics between languages, we decided on using an emotional lexicon that was available in both English and Spanish. One of such corpora is Affective Norms for English Words (ANEW) [1] and its

[2] http://www.geonames.org

Spanish couterpart ANSW [7]. Both of these provide ratings for 1035 words in a valence, arousal and dominance scale. Additionally, they provide with a frequency statistic of how many times a word appeared. Unfortunately, taken as it is, most of these words won't appear in a regular tweet. In order to expand the coverage of our dictionaries we decided on the following method:

1. Learn a domain-specific similarity model between English (Spanish) words.
2. For each word w in ANEW (ANSW):
   (a) Find n similar words in the domain and their similarity ratings.
   (b) For each similar word s with similarity rating $\eta_s > \tau$:
   Asign a valence and arousal rating as follows:

$$valence(s) = valence(w) * \eta_s$$

$$arousal(s) = arousal(w) * \eta_s$$

We assume that similar words must have similar valence and arousal ratings, and that this similarity is mediated by the cosine score. Our domain-specific similarity model was learned using 400 dimension Word2Vec [4] on $12,212,416$ tweets for English and $10,522,314$ for Spanish. We set n to be 100 and the threshold τ to 0.5. The resulting expanded ANEW dictionary had $10,075$ (ANSW, $11,200$) words.

2.2.2 Calculating sentiment of a tweet

In order to obtain a tweet-level sentiment, we did the following: (i) find the words in the tweet that are in the expanded ANEW (ANSW) dictionary; (ii) If the tweet has 1 or less words in the dictionary, classify it as a neutral sentiment; (iii) Else, statistically average the word's valence and arousal. Our statistical average function was a weighted average, using the frequency of a word as the weight of that emotion. Thus a word that was seen 100 times in ANEW carries the double emotion than one only seen 50 times.

3 Results

3.1 By Language

We obtained arousal and valence ratings for all the tweets. One out of every four tweets (24.498%) in English had two or more words with emotion (i.e.,

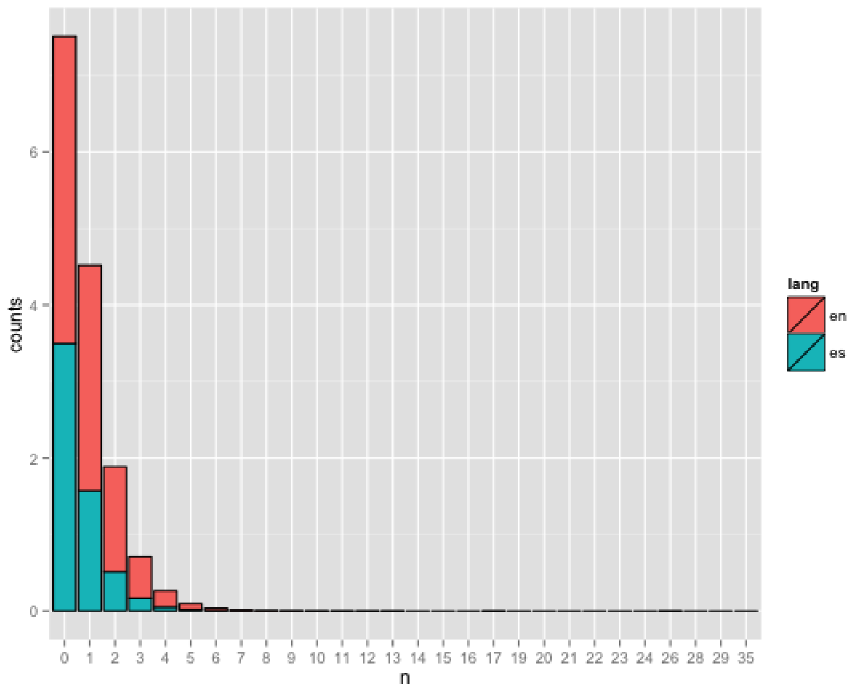

Figure 1: Distribution of words found in ANEW and ANSW

found in the expanded ANEW dictionary). In contrast, only 12.923% had two or more words in the Spanish expanded dictionary. Figure 1 shows the distribution of number of dictionary words found in the tweets according to their language. There was no significant difference in the distributions of emotional words in tweets (KS-test, $p > 0.05$)

3.1.1 By Geography

4 Conclusion

References

[1] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, 1999.

[2] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.

[3] Nikolaos Malandrakis, Michael Falcone, Colin Vaz, Jesse Bisogni, Alexandros Potamianos, and Shrikanth Narayanan. SAIL: Sentiment Analysis using Semantic Similarity and Contrast. *SemEval 2014*, page 512.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[5] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose. *CoRR*, abs/1306.5204, 2013.

[6] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015.*

[7] Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605, 2007.