Using location and dialect to classify Twitter users from the Netherlands and Flanders

Ali Hürriyetoğlu, Piet J.H. Daas

Koç University, Turkey; Statistics Netherlands, Center for Big Data Statistics

INTRODUCTION

- The usefulness of the content on social media for a particular statistical use depends on the characteristics of the users that contribute to generating it.
- Thus, the level of our understanding of the user characteristics affects the quality and the completeness, or selectivity, of the outcome.
- The study reported here focusses on classifying users that write tweets in the Dutch language, in one of its two main dialects:
  1. The language used in the Netherlands (simply called Dutch from hereon),
  2. The language used in Flanders (Flemish).

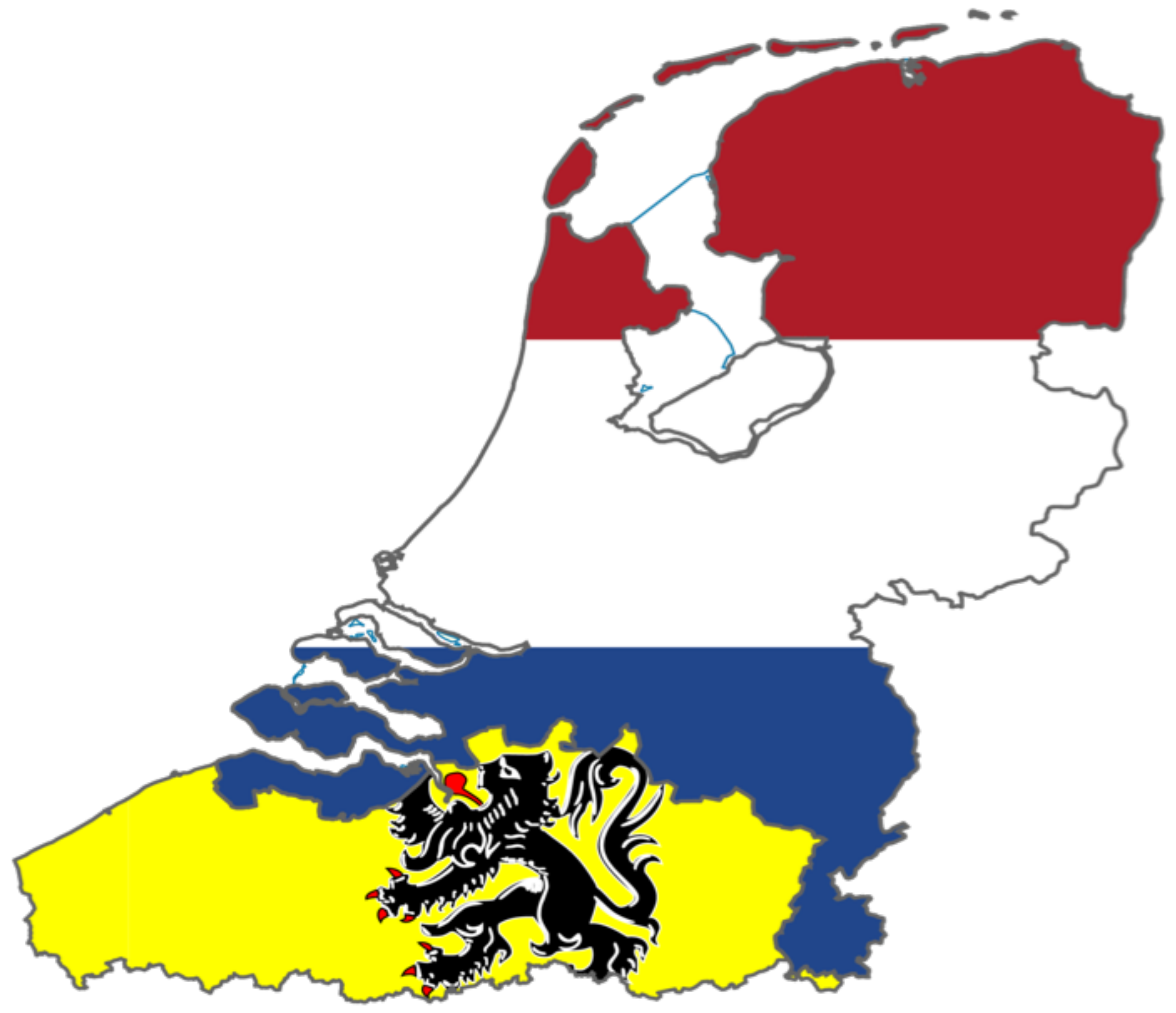

Figure 1 A map of the locations indicating populations speaking Flemish (yellow) and Dutch (red, white, blue)

REFERENCES

[1] GeoNames (2017). http://geonames.org/. Retr. October 5, 2017.
[2] McGuire, P. (2007). Getting started with pyparsing. " O'Reilly Media, Inc.".
[3] Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The twitter of babel: Mapping world languages through microblogging platforms. PloS one, 8(4), e61981.
[4] Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J., Moortgat, M. and Baayen, H. (2002). Experiences from the spoken dutch corpus project, in M. Gonz´alez Rodriguez and C. Paz Su´arez Araujo (eds), Proceedings of the third International Conference on Language Resources and Evaluation, pp. 340–347.
[5] Tjong Kim Sang, E., van den Bosch, A (2013). Dealing with big data: The case of Twitter. Computational Linguistics in the Netherlands Journal , 3:121–134, 12/2013.

Center for Big Data Statistics

USER LOCATION DETECTION

- The location field of all users in the TwiNL [5] database who tweeted at least once in 2016 were studied.
  1. There are a total of 9,130,587 unique ID's.
  2. 5,912,786 has one of the 1,484,427 unique non-empty location fields on their profile.
- The goal was to detect locations in the Netherlands and Belgium from the location field of a user profile.

Language Resources

- A list of location names for Belgium and Netherlands was obtained from Geonames [1].
- Alternative names of the locations from these lists are used as well.
- Location names that were shorter than 3 characters or had one of the common nouns 'brand', 'wel', and 'gem' were excluded due to the high level of ambiguity they entail.
- Place names that occur in both of the countries were handled separately.
- The lists were extended with alternative names and specific abbreviations often used such as "a'dam", "r'dam" and "the hague".

Preprocessing

- The user location strings are pre-processed to remove the most frequent characters that do not contribute to the place name meaning from the beginning and end of a location string.
  - ▶️, ⬅️, ⭕, ♻️, ♊, ♐, ⛵, ⚪, ☕, ♋, ⚫, ✔️, ➡️, ✅, ♦️, ▼

Detection

- Place recognition was performed using Pyparsing (McGuire, 2007).
  - Location = city_name + punc +country_name_NL
  - punc = Lit(',')
  - country_name_NL = oneOf(['NL','Nederland','holland'])
  - City_name = 'Nijmegen'|'Amsterdam'|'020'
- The grammar above recognizes an expression 'Nijmegen, NL'.

Results

- There are 1,484,427 unique location strings in the dataset.
- 54,301 location strings could be parsed as in Netherlands or in Belgium.
- The location strings that were not recognized as a place name were mainly related to
  - places other than the Netherlands or Belgium,
  - used a non-standard name variant,
  - contained a spelling error,
  - or were not a place name at all.

USER CLASSIFICATION

- Almost all of the users use natural language to express themselves.
- The language used reveals the geographical area of a user [3].
- Build a machine learning model that is able to identify the two main groups of Dutch speaking users; those from the Netherlands and those from the upper part of Belgium.

Training Data

- We used 'Corpus Gesproken Netherlands' [4] as training data.
- We downloaded up to 300 tweets from 2,807 and 407 users from the Netherlands and Flanders respectively.

Machine Learning

- Unigram and bigram features were used with Tf-Idf weigthting.
- Support Vector Machines (SVM) classifier trained by optimizing its parameters on 60% of the user tweets downloaded.

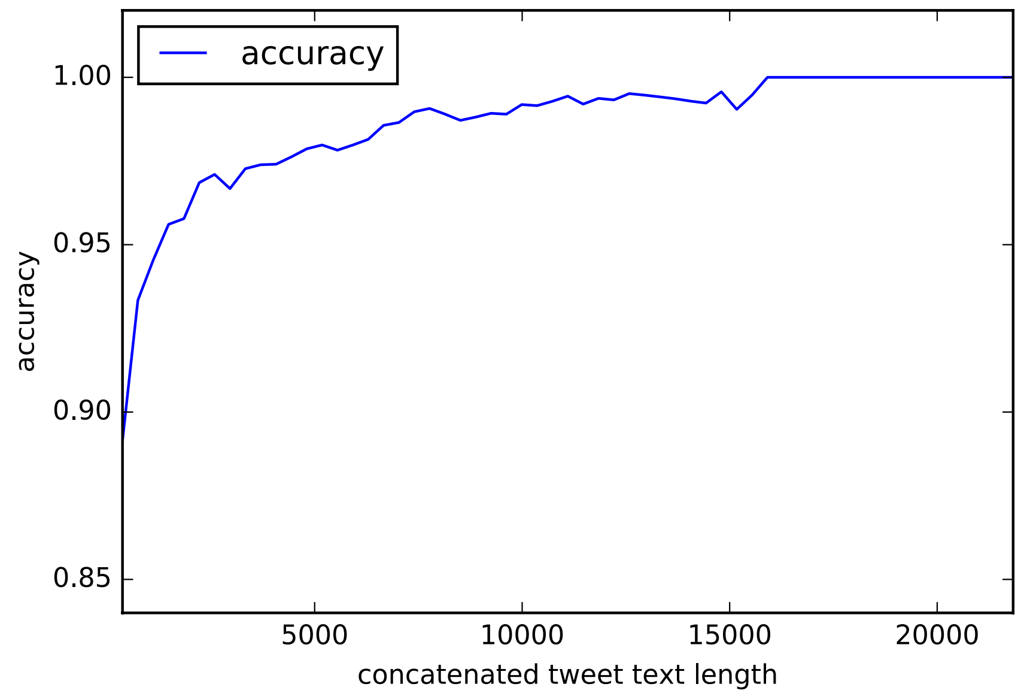

Figure 2 Classifier accuracy dependence on the length of the text (number of tweets) analysed per user

Results

- The classifier yields 98% F1 score on the 40% of the user tweets that were held-out at the training phase.
- As illustrated in Figure 2, the more text we use from a user, the more precise prediction can be made by the classifier.
- Typical words for each class as learned from the training data are:
  - **Netherlands**: nou, Nederland, gulden, Amsterdam, schip, besloten, oh, wanwege, leuk, uh uh, ineens, lekker, Utrecht, KPN
  - **Belgium**: intussen, Vlaamse, Gent, voorbije, men, ge, zei, Frank, he, Vlaanderen, allee, Brussel, Filip, Antwerpen, neen, wellicht, gij