LTRC_IIITH at IBEREVAL 2017: Stance and Gender Detection in Tweets on Catalan Independence

Sahil Swami, **Ankush Khandelwal**, **Manish Shrivastava** and **Syed Sarfaraz Akhtar**

Language Technologies Research Centre, International Institute of Information Technology, Hyderabad

Abstract. We describe the system submitted to IBEREVAL-2017 for stance and gender detection in tweets on Catalan Independence [1]. We developed a supervised system using Support Vector Machines with radial basis function kernel to identify the stance and gender of the tweeter using various character level and word level features. Our system achieves a macro-average of F-score(FAVOR) and F-score(AGAINST) of 0.46 for stance detection in both Spanish and Catalan and an accuracy of 64.85% and 44.59% for Gender detection in Spanish and Catalan respectively.

1 Introduction

The task of opinion mining and sentiment analysis on natural language texts in social media has gained a lot of popularity and importance in recent times. Stance detection is related to sentiment analysis but is very different from it. In sentiment analysis we check if a tweet has a positive, negative or neutral emotion while in stance detection we check whether the tweet is in favor, neutral or against a given target which in this paper is independence of Catalonia. For example, consider the following sentence: "*Recent studies have shown that global warming is in fact real*". We can say that this sentence's author is most likely to be in favor of the concept 'global warming'.

There have been several experiments [2],[10] in the field of sentiment analysis and opinion mining on social media text. Opinion mining can provide a lot of information about the texts that are present in social media and benefits a lot of other tasks such as information retrieval, text summarization, etc.

On the other hand gender detection is the task of inferring the gender of author from the content of the tweet. Gender detection has many applications in the field of marketing and advertising and thus there have been a lot of studies [3],[4],[6],[11] on gender detection in social media text. Twitter profiles don't provide a field for person's gender which makes the task of identifying author's gender from the tweet much more important.

In this paper we present a system for stance and gender detection in tweets. Our system uses character and word level features and Support Vector Machines with radial basis function kernel for classification.

2 Dataset and Evaluation

The organizers provided training and test dataset which consisted of 4319 tweets and 1081 tweets respectively, for both Spanish and Catalan. All the tweets in the training dataset are annotated with stance ('FAVOR' or 'AGAINST' or 'NONE') and gender ('FEMALE' or 'MALE').

Stance detection systems are evaluated using macro-average of F-score (FAVOR) and F-score (AGAINST) i.e.

$$(Fscore_{FAVOR} + Fscore_{AGAINST})/2$$

On the other hand gender detection systems are evaluated using accuracy i.e. number of tweets for which the gender is predicted correctly per hundred tweets.

3 System Framework

3.1 Pre-processing

Initially tweets are tokenized in a way such that hashtags, URLs and mentions are preserved. Then URLs, mentions and stopwords are removed from the tweets.

It can be observed from the tweets present in the training and test datasets that almost all the hashtags are written in camel case format. Therefore, '#' is removed from the hashtags and all the words are extracted from the hashtag. And then each word is considered as a separate token.

All the tokens in Spanish are then stemmed using Snowballstemmer implemented in NLTK.

3.2 Features

We extracted various features from the given tweets to train our machine learning model. We list and describe these features below.

Character N-grams Character n-grams feature refers to presence or absence of contiguous sequence of n characters. It can be seen from previous work [2],[3],[4] that character level features have a significant effect on stance and gender detection.

We extract character n-grams for all values of n between 1 and 3. Including all the n-grams increases the size of feature vector enormously. Therefore, we consider only those n-grams in our feature vector which occur at least 10 times in the training dataset. This reduces the size of feature vector significantly and also removes noisy n-grams.

Word N-grams Word n-grams feature refer to presence or absence of contiguous sequence of n words or tokens. Word n-grams have proven to be important features for stance and gender detection in previous studies [5],[6]. We extract word n-grams for all values of n between 1 and 5. We include only those n-grams in our feature vector which occur at least 10 times in the training dataset.

Stance and Gender Indicative Tokens This feature refers to presence or absence of stance and gender indicative tokens. We use a variation of the approach to find stance indicative hashtags [2] and extract stance and gender indicative tokens. We calculate a score for each token for both stance and gender where score is defined as :

$$Score_{stance}(token) = max_{stance\_label \in Stance-Set} \frac{freq(token, stance\_label)}{freq(token)}$$

$$Score_{gender}(token) = max_{gender\_label \in Gender-Set} \frac{freq(token, gender\_label)}{freq(token)}$$

where Stance-Set = {FAVOR, AGAINST, NEUTRAL}, Gender-Set = {MALE, FEMALE}.

We consider only those tokens as features for stance indication which have a score ≥ 0.6 and occur at least five times in the training dataset. For gender indication we consider only those tokens which have a score ≥ 0.7 and occur at least twice in the training dataset. The threshold value for scores and number of occurrences has been decided after empirical fine tuning.

3.3 Feature Selection

Previous studies [4],[7] have shown that feature selection algorithms improve efficiency and accuracy of classification systems. We use chi square feature selection algorithm which uses chi-squared statistic to evaluate individual feature with respect to each class. This algorithm was run for both stance and gender detection in order to extract the best features and reduce the feature vector size.

3.4 Classification approach

Support Vector Machines have been used many times previously [2],[8],[9] for stance and gender detection and has proven to be a very effective classification technique for the same.

After pre-processing the dataset and extracting all the desired features, we use scikit-learn Support Vector Machine implementation with a radial basis function kernel for classification. We also perform 10-fold cross validation on the provided training dataset to develop the system. 10-fold cross validation is run for each of the individual features separately to observe the effect of each feature on classification.

3.5 Results

Our system achieves a macro-average of F-score(FAVOR) and F-score(AGAINST) of 0.46 for stance detection in both Spanish and Catalan and an accuracy of 64.85% and 44.59% for gender detection in Spanish and Catalan respectively for the given test dataset.

Table 1. shows the accuracy in percentage achieved for stance and gender detection for Spanish tweets while Table 2. shows the accuracy in percentage achieved for stance and gender detection in Catalan tweets considering one feature at a time and also considering all the features together. These are the results achieved in 10-fold cross validation on training dataset.

Table 1. Feature-wise accuracy (in %) for stance and gender detection in Spanish tweets.

	Stance Detection	Gender Detection
Character N-grams	74.94	69.18
Word N-grams	74.03	63.38
Stance and gender indicative tokens	75.40	63.43
All features	75.81	69.83

Table 2. Feature-wise accuracy (in %) for stance and gender detection in Catalan tweets.

	Stance Detection	Gender Detection
Character N-grams	81.16	73.64
Word N-grams	79.48	69.60
Stance and gender indicative tokens	80.64	71.34
All features	81.53	75.38

4 Conclusion and Future Work

In this paper, we presented our approach for stance and gender detection for tweets in both Spanish and Catalan using character and word level features and Support Vector Machine technique for classification. It can also be observed from the results of 10-fold cross validation on training dataset that character n-grams have a significant effect on classification.

Our system has a lot of room for improvement and future work will include extracting more features such as POS-tags and word embeddings and using several other supervised and unsupervised machine learning algorithms for classification.

References

[1] Taulé M., Martí M.A., Rangel F., Rosso P., Bosco C., Patti V. Overview of the task of Stance and Gender Detection in Tweets on Catalan Independence at IBEREVAL 2017. In *Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017.*

[2] Saif M. Mohammad, Parinaz Sobhani, Svetlana Kiritchenko. 2016. Stance and Sentiment in Tweets. In *ACM Transactions on Embedded Computing Systems.* Vol. 0, No. 0, Article 0.

[3] Na Cheng, R. Chandramouli, K.P. Subbalakshmi. 2011. Author gender identification from text. In Digital Investigation. Vol. 8, 78–88.

[4] Zachary Miller, Brian Dickinson, Wei Hu. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. In *International Journal of Intelligence Science* (2012), 2, 143–148.

[5] Peter Krejzl, Barbora Hourová, Josef Steinberger. Stance detection in online discussions. In *CoRR*(2017).

[6] Claudia Peersman, Walter Daelemans, Leona Van Vaerenbergh. 2011. Predicting Age and Gender in Online Social Networks. In *SMUC '11*, Proceedings of the 3rd international workshop on Search and mining user-generated contents (2011). 37–44.

[7] Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, Kenneth Steimel, Sandra Kubler. 2016. IUCL at SemEval-2016 Task 6: An Ensemble Model for Stance Detection in Twitter. In *Proceedings of SemEval*(2016). 394–400.

[8] Clay Fink, Jonathon Kopecky, Maksym Morawski. 2012. Inferring Gender from the Content of Tweets: A Region Specific Example. In *Sixth International AAAI Conference on Weblogs and Social Media*(2012).

[9] James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, Martine De Cock. 2014. Age and Gender Identification in Social Media. In *CLEF 2014 Working Notes proceedings.*

[10] Antonio Fern andez Anta, Luis Nune Chiroque, Philippe Morere, Agustın Santo. 2013. Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of NLP Techniques. In *Procesamiento de Lenguaje Natural*(2013). 50:45–52.

[11] John D. Burger, John Henderson, George Kim, Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* . 1301—1309.