Dimensionality Reduction of Unstructured and Network Data for Stance Detection

Andrea Sciandra[1]

[1]University of Modena and Reggio Emilia – andrea.sciandra@unimore.it

Abstract

The idea behind this work stems from the participation in some shared tasks concerning stance detection in NLP conferences. In these competitions, participants tried to develop the best stance prediction system for 'favor', 'against', and 'none' categories on selected topics, according to messages and relationships among users of a social networking site. Thus, the data available consisted of textual and network data. The teams we collaborated with used dimensionality reduction methods for network data, through a Multidimensional Scaling. On the other hand, the approach towards textual data involved different methods of feature extraction, without paying particular attention to dimensionality reduction for unstructured data. In this paper we show the empirical results of a two-step strategy to obtain lower-dimensional textual data relying on text mining techniques and principal component analysis. The results show levels of accuracy comparable to classical feature extraction techniques and to the best task models, despite using a much smaller number of predictors.

Keywords: Dimensionality Reduction, Stance Detection, Social Media, PCA, MDS, Multinomial Regression.

1. Introduction

Stance detection is the task of automatically classifying whether the author of a text is in favor, against, or neutral towards a given target. The target could be a person, an organization, a government policy, a movement, a product, etc. (Mohammad et al., 2017). Stance detection could be useful for several undertakings, such as market analysis, opinion surveys, predictions for elections, policy-making, media monitoring, and security (Küçük & Can, 2020). In fact, a stance detection could support the automatic identification of people's extremist tendencies on the one hand, but it could also be employed by authoritarian governments as a tool to control their citizens (Lai et al., 2021).

The idea behind this work stems from our participation in some shared tasks concerning stance detection in NLP conferences (Evalita 2020[1] and IberLEF 2021[2]). In these competitions, participants tried to develop the best stance prediction system for 'favor', 'against', and 'none' categories on selected topics, according to messages and relationships among users of a social networking site (Twitter). The shared tasks dealt with the stance detection of Italian tweets about the Sardines movement (SardiStance) and Spanish or Basque tweets about the Antivaxxers movement (VaxxStance). In particular, SardiStance was the first shared task

[1] www.evalita.it/evalita-2020

[2] sites.google.com/view/iberlef2021

regarding Stance Detection in Italian tweets, where the organizers collected posts about the Sardines movement and invite at automatically detecting their stance (Cignarella et al., 2020).

The data available consisted primarily of textual and network data. In both cases, the teams we collaborated with used dimensionality reduction methods for network data, through a Multidimensional Scaling (MDS) applied to distance matrices among social media users. On the other hand, the approach towards textual data involved different methods of feature extraction, without paying particular attention to dimensionality reduction for unstructured data. In this paper we show, regarding the SardiStance tasks, the empirical results of a two-step strategy to obtain lower-dimensional textual data relying on text mining techniques and principal component analysis (PCA) of the reduced document term matrix (DTM).

2. Data processing

The data available for these competitions included the text of the posts, information about the posts (number of retweets, number of favors, etc.), information about the authors (number of followers, number of friends, etc.), and information about their social networks (friends, replies, retweets, and quotes' relationships). In this contribution we focused particularly on post texts and social networks of friendship and retweeting, therefore dealing essentially with unstructured data and network data.

Our strategy for limiting the dimensionality of unstructured data is based on two steps. The first step involves treating the textual corpus through text mining techniques for features reduction. Specifically, the processes of text normalization, lemmatization, less frequent terms removal, and normalized TF-IDF calculation were applied. The text normalization included recognition of URLs and emoticons, stopwords removal, and slang/jargon correction through the `TextWiller` R package (Solari et al., 2019). Following normalization and lemmatization, lemmas with fewer than 30 occurrences were removed. In this way, we reduced the initial corpus containing 14815 types and 89183 tokens to a reduced corpus containing 254 types and 23338 tokens.

The second step involves a dimensionality reduction through statistical techniques. When we have data with a large number of dimensions, one goal may be to project this data into a lower dimensional subspace, while trying not to lose valuable information about the original variables. One way to achieve this reduction is through the selection of variables, also called feature selection. Another way is through the creation of a reduced set of linear transformations of the initial variables. The creation of these composite features, using projection techniques, is often referred to as feature extraction. Principal component analysis (PCA) is a technique designed to derive, from a set of correlated numerical variables, a smaller set of orthogonal variables. The reduced set of linear orthogonal projections is obtained by linearly combining the original variables. Based on these considerations, we extracted the principal components from the reduced DTM, by first standardizing the normalized TF-IDFs. Following the PCA, the first 100 resulting dimensions were selected (cumulative proportion of variance: 50.7%) and used as features to predict the stances.

Conversely, with respect to network data, our approach to reduce the dimensionality of the data was based on the MDS applied to distance matrices among social media users (minimum path of graphs of 'friendship' ties and ties generated by retweets). In this work we consider only friendship and retweeting relationships because reply and quote graphs showed few relationships. So, for each network, a distance matrix among Italian twitterers was computed.

We defined the distance among two users as the shortest path, forcing the graph to be undirected. The distance matrix was then projected into a Euclidean space through a MDS (Klimenta & Brandes, 2012). We expected the users to be partially polarized in clusters within the networks and, consequently, the largest dimensions of the MDS should polarize the stances. An exploratory data analysis confirmed our expectation: for instance, in Figure 1 (top) we show the scatter plot of dimensions 1-2 and 1-3 for the Friends Network. The first dimension clearly discriminates the three stances, in particular 'favor' and 'against'. Therefore, we decided to retain the four main dimensions for each of the two networks. Finally, we added the 8 dimensions extracted from the MDS coordinates for friendship and retweeting networks to PCA features.

Inspecting the bottom part of Figure 1, we must note that the first dimensions derived from PCA do not seem to discriminate efficiently among stances. Anyway, we expected a similar result due to the unstructured data, but we also expected that the exploitation of several components would be helpful to classify stances with substantial accuracy.

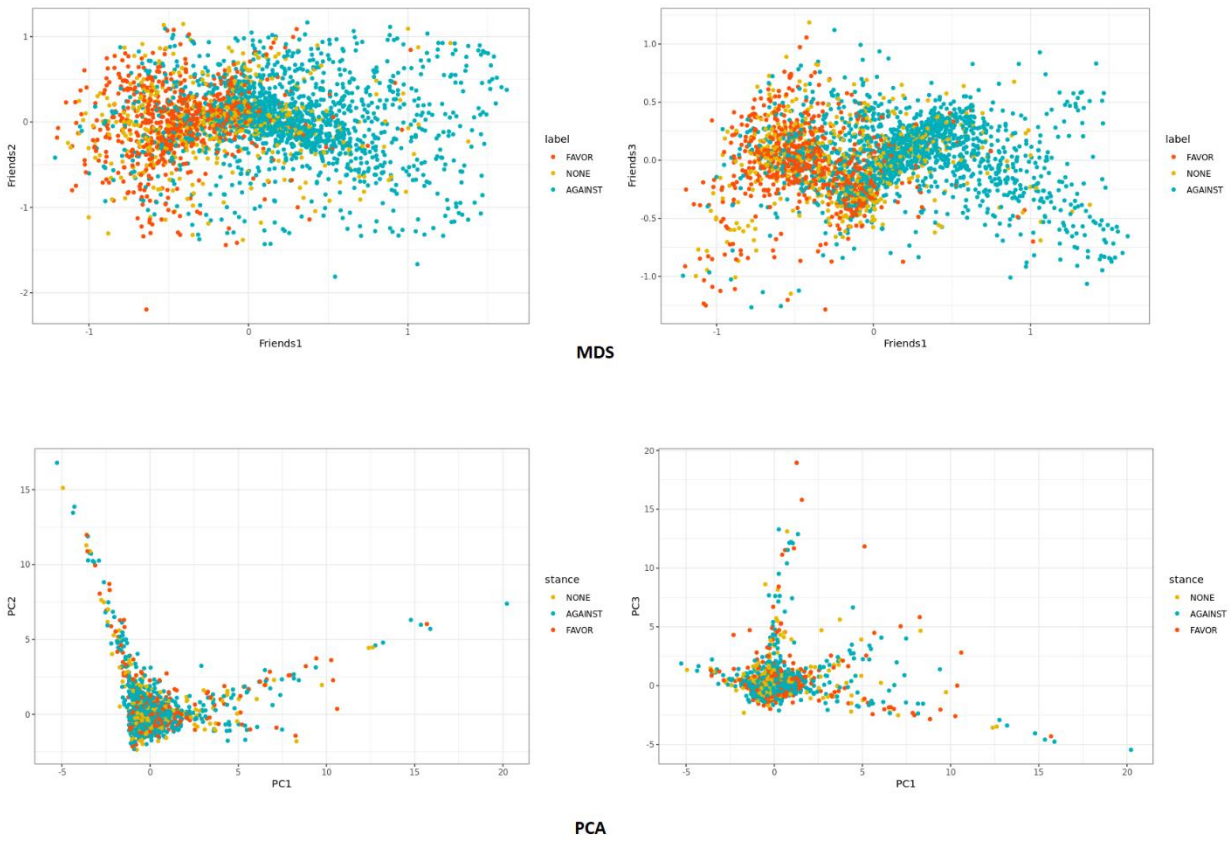

Fig. 1. Scatter plots of the First-Second dimensions and First-Third dimensions extracted by the MDS from the distance matrix of the Friends Network and by the PCA of the DTM.

3. Analysis

3.1. Experiments

The SardiStance competition included two tasks: in task A participants had to predict the stance towards the movement Sardine exploiting only textual information, while in task B teams could

employ contextual information based on: the post, the user, and his/her social networks. The SardiStance organizers then provided a training set (N=2132), a test set (N=1110) and separate files containing information on tweets and users, as well as edge lists for the different relationships between users. Predictive systems were evaluated using the F1-score computed over the two main classes (favor and against). For each task, organizers also computed a baseline using a simple statistical learning model based on SVM combining uni-gram features.

In this paper we show the results of three experiments: the first experiment concerns the prediction of stances using only the first 100 principal components of the DTM (task A), the second adds to the features the 8 dimensions derived from the MDS (task B), the third also adds some context information (task B), such as the number of: posts, friends, and followers for each user; listings, retweets, and favorites for each tweet.

We chose as statistical learning model a penalized multinomial logistic regression, using a 5-fold cross-validation (with 100 repetitions) tuning of the decay parameter through the `nnet` R package. Therefore, we used a single-hidden-layer neural network, where the aim of the decay parameter is to prevent the weights to be too large through penalizations during the fitting process. This procedure should help avoiding overfitting (Venables et al., 2013). We chose this classifier because, even though it does not need the independent variables to be statistically independent from each other (unlike other classifiers, e.g., Naïve Bayes), it assumes collinearity to be low (Belsey, 1991).

3.2. Results

Table 1 shows for each experiment the global level of accuracy, the F1-score for the stances 'against' and 'favor', and their F1-average. Measures were computed on the labeled (gold) test set. The last column shows the ranking position in which each model would have been placed if it had participated in the relative task of the competition. The results of the application of these techniques show levels of accuracy comparable to classical feature extraction techniques and to the best task models, despite using potentially a much smaller number of predictors, and without considering n-grams. Moreover, our models for task B did not include other information on the tweet itself (source, creation date, etc.), contextual information about the user (creation date, emoji in bio, etc.), quotes and replies networks.

Table 1. Results of the experiments (test set – gold labels) using 100 principal components of the DTM (PC), 8 dimensions extracted from MDS coordinates (MDS), and contextual information (context).

Experiment	Accuracy	F1-against	F1-favor	F1-Avg (against, favor)	Rank in task[3]
1 (PC)	0.6099	0.7506	0.3325	0.5415	16th
2 (PC,MDS)	0.7189	0.8440	0.5950	0.7195	5th
3 (PC,MDS,context)	0.7180	0.8426	0.6095	0.7261	3rd

[3] www.di.unito.it/~tutreeb/sardistance-evalita2020/index.html

In the SardiStance competition, our team, TextWiller (Ferraccioli et al., 2020), also used 50 vectors derived from word embedding as features, improving F1-average of Experiment 3 by a very small amount (+0.0048). This result does not particularly surprise us, because we have already found that using word embeddings as predictive features does not improve the results compared to frequencies (Sciandra, 2020) or TF-IDFs of words, as in this work. We believe that a smarter use of word embeddings, e.g., by means of the dependency relations between the target of interest (Sardines here) and the connected tokens in the dependency tree (Lai et al., 2021), can lead to better results. Instead, the inclusion of information about social networks increased the F1-average remarkably (+0.18). In fact, the Experiment 1 with only the principal components of the DTM ranks below the baseline of about 0.037 points of the F1-average, whereas Experiment 2 ranks 0.09 points above the baseline of Task B. Moreover, the analysis of the most important variables[4] in Experiment 3 also shows the dominance of network measures (Fig. 2).

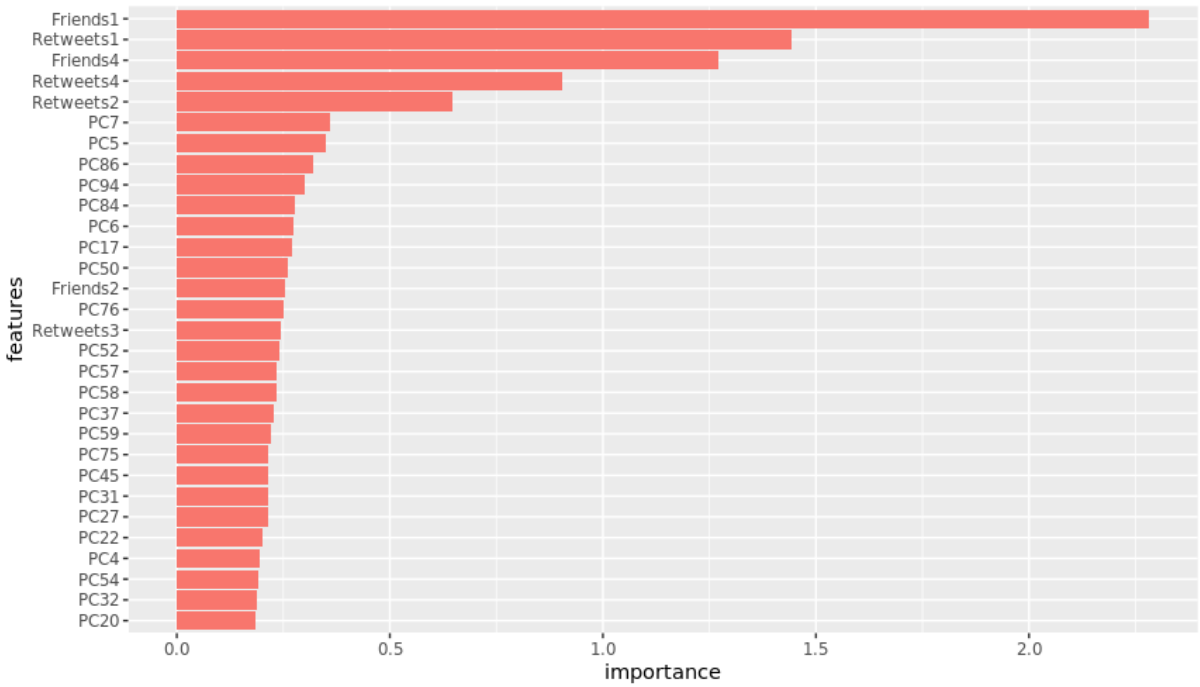

Fig. 2. Variable importance (top 30) – Experiment 3

Specific tests on each variable, such as Wald's test or Likelihood Ratio test, may be useful to complete the analysis, along with the relative risks (odds ratios).

4. Conclusion

The main advantage of this approach, in our opinion, lies in the possibility to extract a reduced number of orthogonal dimensions that maximize the ability to capture the variability of textual features. The approach of this work proved to be also highly competitive with respect to dimensionality reduction techniques related to the semantic scope, such as word embedding.

[4] The method of computing the importance of variables in a neural network exploits combinations of the absolute values of the weights (Gevrey et al., 2013).

Clearly, we need to improve the performance of models based only on PCA. For instance, more elaborate state of the art methods of feature extraction for texts (e.g., n-grams, stylistic, lexicon-based, and dependency-based features) surely can enhance the prediction accuracy over our approach, but they often result in very sparse and/or highly collinear matrices, leading to possible computational problems and overfitting risk.

Finally, we point out that reducing the dimensionality of the ties among users showed great efficiency, significantly improving the performance of the prediction model with an extremely limited number of dimensions.

Acknowledgements

This research has been partially supported by the University fund for Research (FAR 2020) of the University of Modena and Reggio Emilia.

References

Belsley D. (1991). *Conditioning diagnostics: collinearity and weak data in regression*. New York: Wiley.

Cignarella A. T., Lai M., Bosco C., Patti V. and Rosso P. (2020). SardiStance@ EVALITA2020: Overview of the task on stance detection in Italian tweets. In *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* (pp. 1-10). CEUR Workshop Proceedings.

Ferraccioli F., Sciandra A., Da Pont M., Girardi P., Solari D. and Finos L. (2020). TextWiller@ SardiStance, HaSpeede2: Text or Con-text? A smart use of social network data in predicting polarization. *Proc. of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020).* CEUR Workshop Proceedings.

Gevrey M., Dimopoulos I. and Lek S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3): 249-264.

Klimenta M. and Brandes U. (2012). Graph drawing by classical multidimensional scaling: new perspectives. In *International Symposium on Graph Drawing*. Springer, Berlin, Heidelberg, pp. 55-66.

Küçük D. and Can F. (2020). Stance detection: A survey. *ACM Computing Surveys (CSUR)*, *53*(1): 1-37.

Lai M., Cignarella A. T., Finos L. and Sciandra A. (2021). WordUp! at VaxxStance 2021: Combining Contextual Information with Textual and Dependency-Based Syntactic Features for Stance Detection. In *XXXVII International Conference of the Spanish Society for Natural Language Processing.* (Vol. 2943). CEUR Workshop Proceedings, pp. 210-232.

Mohammad S. M., Sobhani P. and Kiritchenko S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, *17*(3): 1-23.

Sciandra A. (2020). COVID-19 Outbreak through Tweeters' Words: Monitoring Italian Social Media Communication about COVID-19 with Text Mining and Word Embeddings. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, pp. 1-6.

Solari D., Sciandra A. and Finos L. (2019). TextWiller: Collection of functions for text mining, specially devoted to the Italian language. *Journal of Open Source Softwar*e, 4(41): 1256.

Venables W. N. and Ripley B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.