SINAI at SemEval-2017 Task 4: User based classification

Salud María Jiménez-Zafra, Arturo Montejo-Ráez,
M. Teresa Martín-Valdivia, L. Alfonso Ureña-López
Computer Science Department, Escuela Politécnica Superior de Jaén
Universidad de Jaén, 23071 - Jaén (Spain)
{sjzafra, amontejo, maite, laurena}@ujaen.es

Abstract

This document describes our participation in SemEval-2017 Task 4: Sentiment Analysis in Twitter. We have only reported results for subtask B - English, determining the polarity towards a topic on a two point scale (positive or negative sentiment). Our main contribution is the integration of user information in the classification process. A SVM model is trained with Word2Vec vectors from user's tweets extracted from his timeline. The obtained results show that user-specific classifiers trained on tweets from user timeline can introduce noise as they are error prone because they are classified by an imperfect system. This encourages us to explore further integration of user information for author-based Sentiment Analysis.

1 Introduction

Task 4 of SemEval 2017, Sentiment Analysis in Twitter (Rosenthal et al., 2017), has included some new subtasks this year. One of these subtasks considers user information to be also integrated in proposed systems. We have participated in subtask B consisting of, given a message and a topic, classify the message on a two-point scale (positive or negative sentiment towards that topic). Actually, organizers provide scripts to download user profile information such as age, location, followers... We have taken advantage of this information to expand a SVM model trained with Word2Vec vectors from user publications on this social media.

In this paper, we present our approach to classify tweets in a two point scale (positive and negative) by combining Support Vector Machine (SVM), Word2Vec (Mikolov et al., 2013) and user information. We have decided to combine these technologies for several reasons. Firstly, we have applied SVM many different tasks including tweet polarity classification with good results (Saleh et al., 2011). Secondly, after a revision of the systems presented in the last year for the same task (Nakov et al., 2016), it seems that better results are achieved by using word embeddings representations, so we have decided to test how it works on user modeling. Finally, this year for the first time, organizers include user information. We consider that it is very interesting to integrate this contextual information to improve tweets sentiment classification. Actually, polarity classification on a per-user basis has been found to be useful in tasks like collaborative filtering (García-Cumbreras et al., 2013). Besides, the generation of user profiles in Twitter has attracted the attention of many researches in recent years, enabling the prediction of user behavior as in election processes (Pennacchiotti and Popescu, 2011).

In Section 2 we explain the data used in our approach. Section 3 presents the system description. Experiments and results are expounded in Section 4 and they are analyzed in Section 5. Finally, in Section 6, conclusions and future work are commented.

2 Data

The organizers provided English data from previous years (2015 and 2016). The test set corresponding to 2016 was also supplied for development purposes but, since then, it can be used for training too. In the experimentation phase, the training set is composed by the development, training and test datasets of 2015 and the development and training datasets of 2016. For our participation in task 4 we used all this data for training. In Table 1, it can be seen the distribution of tweets used in the experimentation and testing phases.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 634–639,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

Set	Positive	Negative	Total
training_dev	6,739	1,674	8,413
dev	8,212	2,339	10,551
training_test	14,951	4,013	18,964
test	2,463	3,722	6,185

Table 1: Number of tweets provided for experimentation and testing.

3 System description

The system presented is based on user modeling. It determines the user opinion on a tweet according to a user model generated from his timeline. In our experiments, all tweets are vectorized using Word2Vec. First, a general SVM model on training vectors is generated. Then, for each user in the test set, the system downloads the last 200 tweets published by the user and classifies them using a general SVM classifier, the one resulting from the training set. If the classified tweets from the timeline contains positive and negative tweets and an specific SVM model of the timeline reports an accuracy over 0.7 on leave-one-out cross-validation, the user model is applied on authored tweets from the test set; if not, the general SVM model is applied. Thus, we try to train a per-user classifier, whenever feasible.

For the Word2Vec representation of the tweets, it has been used the software[1] developed by the authors of the method (Mikolov et al., 2013). In order to get representative vectors for each word, it is needed to generate a model from a large text volume. To this end, a Wikipedia[2] dump in English of the articles in XML was downloaded, and the text from them was extracted. The parameters used have been those that provided better results in previous experiments with Spanish tweets (Montejo-Ráez and Dıaz-Galiano, 2016; Montejo-Ráez et al., 2014): a window of 5 terms, the CBOW model and a number of dimensions expected of 300. In this way, each tweet of the training and test set has been represented with the resultant vector of calculating the average and standard deviation of the Word2Vec vectors from words in the tweet text, resulting in a final vector of 600 features. Previously, a simple normalization has been performed on each tweet: repeated letters have been eliminated, stop words have been removed and all words have been transformed to lowercase.

The SVM implementation selected is that based on LibSVM (Chang and Lin, 2011) provided by the Scikit-learn library (Pedregosa et al., 2011).

4 Experiments and results

Three different experiments were conducted over the development set as follows (Fig. 1 and Fig. 2):

- Experiment 1: a general SVM model on Word2Vec representations of training tweets was generated. Each tweet of the development set was vectorized using Word2Vec and classified with the model obtained previously.
- Experiment 2: each tweet vector was expanded with a user vector. A general SVM model was also generated, but on both the Word2Vec representation of the training tweets and user timeline. For every user in the training tweets, the last 200 tweets from his timeline were downloaded. These tweets were used to enrich the vector of each individual tweet. Each tweet of the development set along with user timeline who posted it were vectorized using Word2Vec and the tweet was classified with the model.
- Experiment 3: the general SVM model of experiment 1 was used but one model per user was also defined. In order to define the user model, the last 200 tweets published by the user were retrieved and each of them was vectorized and classified using the general SVM model. Each tweet of the development set was vectorized using Word2Vec and classified according to the following approach: if the model corresponding to the user contains positive and negative tweets and the leave-one-out cross-validation reports an accuracy over 0.7%, the tweet is classified with the user model; if not, it is classified with the general SVM model.

The results obtained in the development phase are shown in Table 2. Although experiment 1 was the one that provided the best results, for our participation in the task, we selected the approach developed in experiment 3 because it takes into account user information, one of the challenges of this year. Experiment 2 also considers user information and got better results than experiment

[1] https://code.google.com/p/Word2Vec/

[2] https://dumps.wikimedia.org/enwiki/

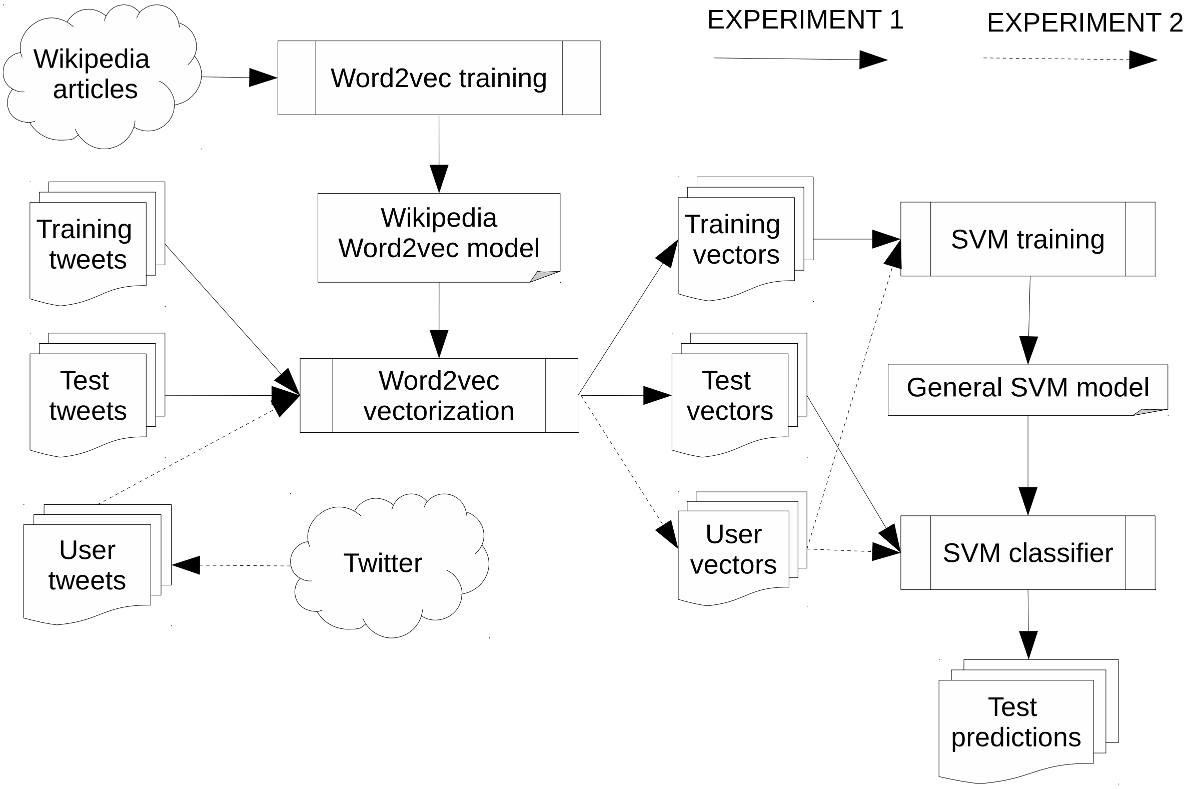

Figure 1: Data flow for experiment 1 and 2.

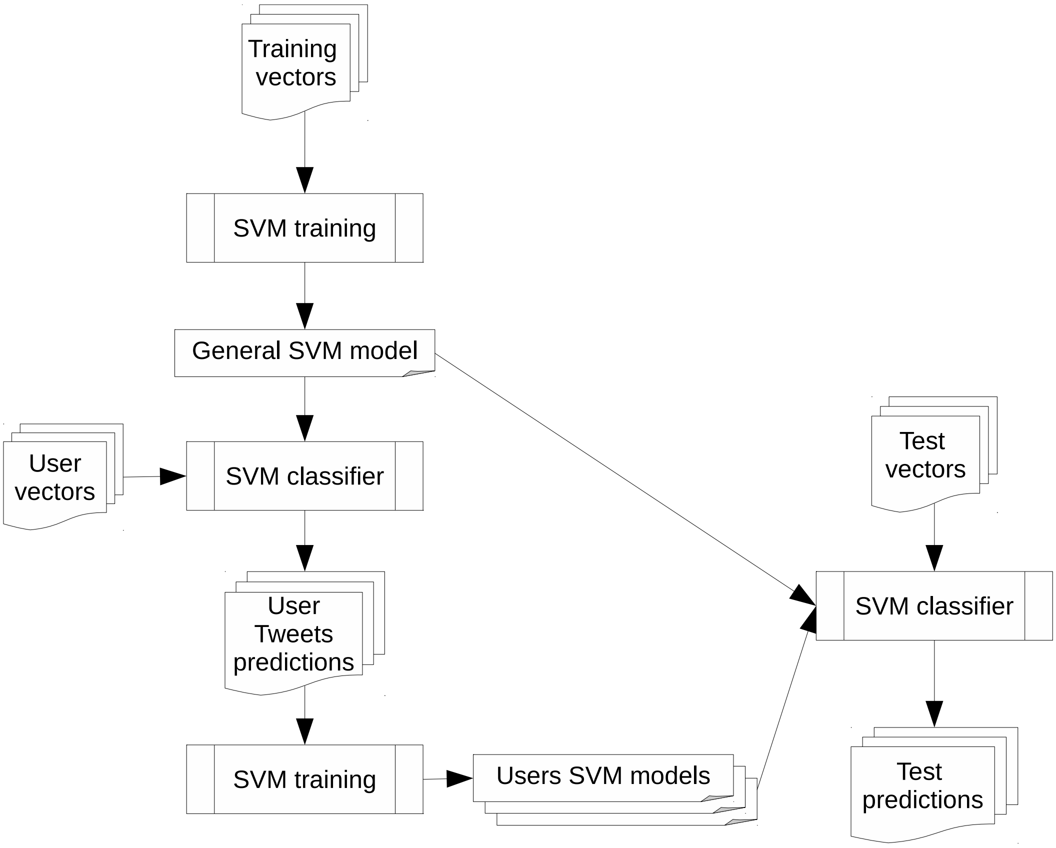

Figure 2: Data flow for experiment 3.

	Exp_1	Exp_2	Exp_3
P positive	0.856	0.854	0.842
P negative	0.764	0.757	0.772
R positive	0.962	0.962	0.970
R negative	0.432	0.422	0.363
Avg. F1	0.729	0.723	0.698
Avg. R	0.697	0.692	0.666
Acc.	0.845	0.842	0.835

Table 2: Results for the development phase.

#	**System**	**AvgR**	**AvgF1**	**Acc**
1	BB_twtr	$\mathbf{0.882}_{1}$	0.890_{1}	0.897_{1}
2	DataStories	$\mathbf{0.856}_{2}$	0.861_{2}	0.869_{2}
3	Tweester	$\mathbf{0.854}_{3}$	0.856_{3}	0.863_{3}
4	TopicThunder	$\mathbf{0.846}_{4}$	0.847_{4}	0.854_{4}
5	TakeLab	$\mathbf{0.845}_{5}$	0.836_{5}	0.840_{6}
6	funSentiment	$\mathbf{0.834}_{6}$	0.824_{8}	0.827_{8}
7	YNU-HPCC	$\mathbf{0.834}_{6}$	0.816_{10}	0.818_{10}
8	WarwickDCS	$\mathbf{0.829}_{8}$	0.834_{6}	0.843_{5}
9	CrystalNest	$\mathbf{0.827}_{9}$	0.822_{9}	0.827_{8}
10	zhangweida2080	$\mathbf{0.826}_{10}$	0.830_{7}	0.838_{7}
11	Amobee-C-137	$\mathbf{0.822}_{11}$	0.801_{12}	0.802_{12}
12	**SINAI**	$\mathbf{0.818}_{12}$	$\mathbf{0.806}_{11}$	$\mathbf{0.809}_{11}$
13	NRU-HSE	$\mathbf{0.798}_{13}$	0.787_{13}	0.790_{13}
14	EICA	$\mathbf{0.790}_{14}$	0.775_{14}	0.777_{16}
15	OMAM	$\mathbf{0.779}_{15}$	0.762_{17}	0.764_{17}
16	NileTMRG	$\mathbf{0.769}_{16}$	0.774_{15}	0.789_{15}
17	ELiRF-UPV	$\mathbf{0.766}_{17}$	0.773_{16}	0.790_{13}
18	DUTH	$\mathbf{0.663}_{18}$	0.600_{18}	0.607_{18}
19	ej-za-2017	$\mathbf{0.594}_{19}$	0.486_{21}	0.518_{19}
20	SSN_MLRG1	$\mathbf{0.586}_{20}$	0.494_{20}	0.518_{19}
21	YNU-1510	$\mathbf{0.516}_{21}$	0.499_{19}	0.499_{21}
22	TM-Gist	$\mathbf{0.499}_{22}$	0.428_{22}	0.444_{22}
23	SSK_JNTUH	$\mathbf{0.483}_{23}$	0.372_{23}	0.412_{23}
	baseline 1: all POSITIVE	0.500	0.285	0.398
	baseline 2: all NEGATIVE	0.500	0.376	0.602

Table 3: Results for SemEval-2017 Task 4, subtask B - English.

3 in the development phase, but we did not select it because we considered that the fact of adding tweets without more sense was not a good idea. Experiment 3 makes more sense, since it defines a personal model for each user based on the way he thinks.

The results for all participants in the test phase can be seen in Table 3 and the detailed report of the results for all participants can be found at (Rosenthal et al., 2017).

Once the gold standard corresponding to the test phase was released, we also conducted other experiments that we defined in the development phase. The results related to the test set in all the experiments are shown in Table 4. Following, in the next section, an in-depth analysis of the results obtained is performed.

	Exp_1	Exp_2	Exp_3
P positive	0.735	0.730	0.718
P negative	0.897	0.890	0.893
R positive	0.862	0.851	0.859
R negative	0.794	0.791	0.777
Avg. F1	0.818	0.812	0.806
Avg. R	0.828	0.821	0.818
Acc.	0.821	0.815	0.809

Table 4: Results for the test phase.

5 Analysis of results

The results obtained do not seem to support the integration of content from users' timelines. In Table 4 we can see that using word embeddings in tweet words straightforward yielded the best results. Adding further user information did not improve the first setup. A model of the user under the form of an aggregated vector computed from his timeline, or a specific polarity classifier for each user involves, first, to download hundreds of tweets for every single user in the data set and, second, use these tweets to compute a final user model.

It is important to note that the SemEval data set is very unbalanced, and that can affect the generation of user classifiers. Besides, not additional data has been used to determine the polarity of tweets in the timeline, so the effects of a bad performance might be, therefore, amplified. Anyhow, experiment 3 shows similar results as the other two approaches, despite the potential bias that recent tweets from the timeline may have on the classification process.

6 Conclusion

Working on timelines has been found interesting as a source of information to generate user profiles (Bollen et al., 2011). Actually, as more text is obtained, further analysis on user behavior or personality can be performed (Diakopoulos and Shamma, 2010).

We will continue exploring how the timeline could be better integrated or analyzed for an effective user modeling process. As the timeline is provided on recent tweets, it could be worth downloading those closer to the moment when the tweet to analyze was published, so the context would be more coherent.

Acknowledgments

This work has been partially supported by a Grant from the Ministerio de Educación, Cultura y Deporte (MECD - scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Ministerio de Economía y Competitividad.

References

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM* 11:450–453.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27.

Nicholas A. Diakopoulos and David A. Shamma. 2010. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, CHI '10, pages 1195–1198. https://doi.org/10.1145/1753326.1753504.

Miguel Á. García-Cumbreras, Arturo Montejo-Ráez, and Manuel C. Díaz-Galiano. 2013. Pessimists and optimists: Improving collaborative filtering through sentiment analysis. *Expert Systems with Applications* 40(17):6758 – 6765.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

A Montejo-Ráez and MC Dıaz-Galiano. 2016. Participación de sinai en tass 2016. *Comité organizador* page 41.

Arturo Montejo-Ráez, MA García-Cumbreras, and M Carlos Díaz-Galiano. 2014. Participación de sinai word2vec en tass 2014. In *Proceedings of the TASS workshop at SEPLN*.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. *Proceedings of SemEval* pages 1–18.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(Oct):2825–2830.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. *Icwsm* 11(1):281–288.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Vancouver, Canada, SemEval '17.

M Rushdi Saleh, Maria Teresa Martín-Valdivia, Arturo Montejo-Ráez, and LA Ureña-López. 2011.

Experiments with svm to classify opinions in different domains. *Expert Systems with Applications* 38(12):14799–14804.