Universidade Vigo

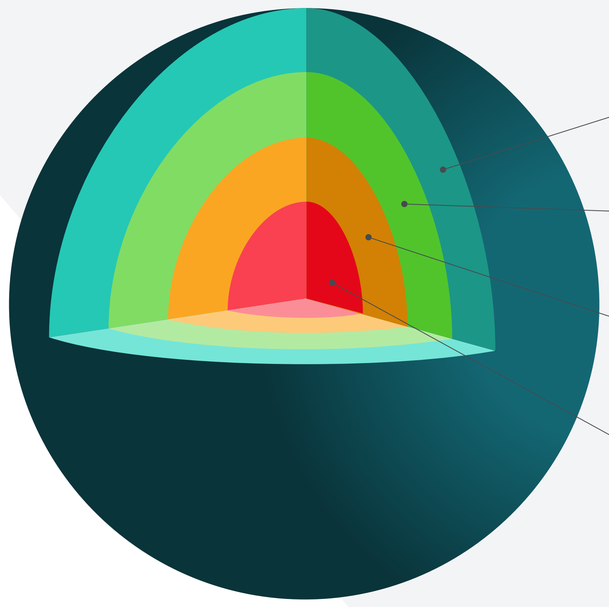
GENERATION OF KNOWLEDGE IN DYNAMIC FINANCIAL MARKETS WITH SUPERVISION

AUTOR: Óscar Barba-Seara

ADVISORS: Milagros Fernández-Gavilanes, Javier González-Castaño

AFFILIATION: AtlanTTIC Research Center, University of Vigo

MOTIVATION OF THE WORK



- Obtain a multi-context solution for short text classification on the financial area
- Test the results on real business context.
- Classify banking movements for personalized marketing according to the user profile/interests
- Make a efficiency & scalable approach to an opportunity in PSD2 environment.
- Great amount of information available online.
- Growth financial solutions and their information.

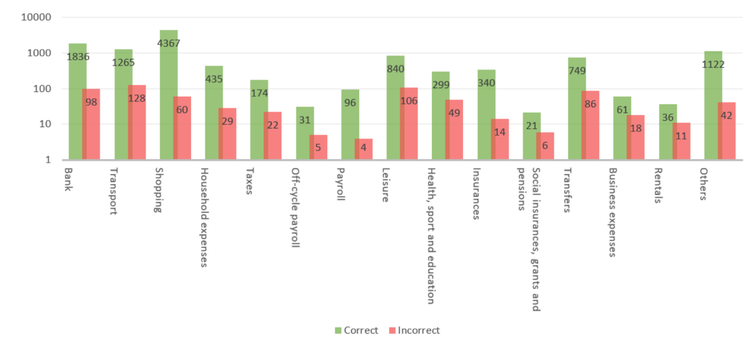
THESIS OBJETIVES

- Application and adaptation to different datasets and contexts.
- Design an automatic SVM system competitive with other existing ones.
- Select the best features to train the SVM classifier
- Compare the results with another SVM & platform results

*** Need to improve the understanding of short texts in financial contents*

RESULTS

- Short texts have less features and higher irregularity than longer texts.
- Short text classification based on a svm classifier combined with linguistic knowledge is used to learn and label financial short text samples. The experimental results show that this method does improve the classification effect of short text.
- Improving efficiency of financial short text classification from available data is still challenging.
- The proposed approach produces better classification accuracy results when lexica knowledge is used as a feature as well as the information related to the amount and date of the banking movement.



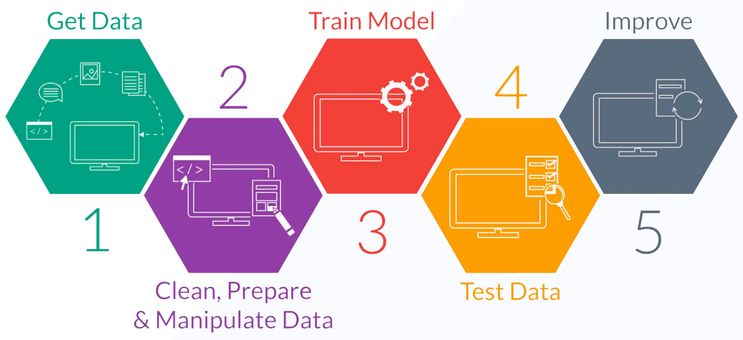
- The proposed framework is composed of three modules:
  - Pre-processing of short text
  - Learning of probabilistic model on probability matrices
  - Classification of banking movements by using the learned model.
- Test on a manual annotated dataset with over 30,871 banking transaction descriptions.
  - Joint Project between GTI & Coinscrap Finance SL [9/2017 and 2/2018]
- Our effectiveness versus other approaches on short text classification

%Train	%Test	Enabled features	P_{macro}	R_{macro}	F_{macro}
70%	30%	Word m-grams	80.76%	50.48%	62.13%
		Word m-grams + amount + date	88.27%	53.79%	66.85%
		Word m-grams + amount + date + lexica	94.63%	81.62%	87.65%
		Word m-grams + amount + date + lexica + char n-gram	95.69%	89.49%	92.48%
		All-In-1b (Plank, 2017)	94.21%	92.16%	93.14%
		IITP-CNN (Gupta et al., 2017)	86.07%	78.43%	79.47%
		IITP-CNN+RNN (Gupta et al., 2017)	93.95%	85.70%	89.14%

MODEL

- Our model is based on a traditional SVM, character and word n-grams as well as linguistic knowledge and other features.
- Real Spanish banking dataset containing fifteen different classes

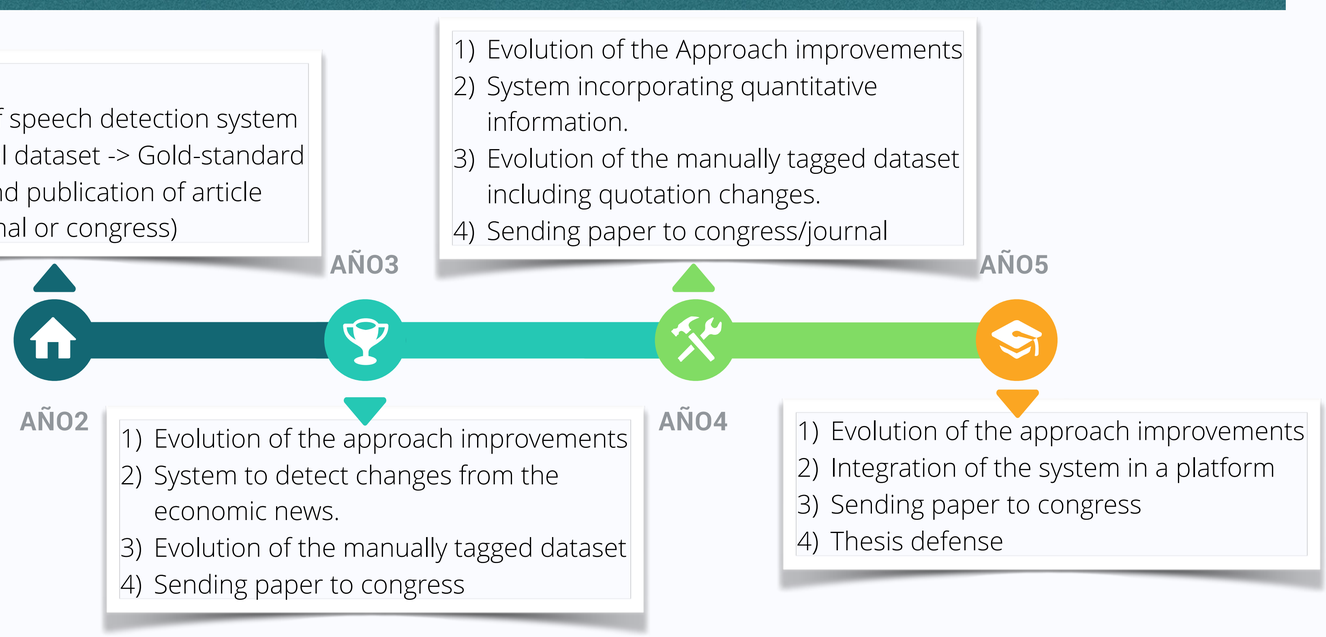
Id	Category
1	Bank
2	Means of transport
3	Shopping
4	Household expenses
5	Taxes and charges
6	Off-cycle payroll
7	Payroll
8	Leisure
9	Health-sport and education
10	Insurances
11	Social insurances, grants and pensions
12	Transfers
13	Business and professional expenses
14	Rentals
15	Others



REFERENCES

- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12, 2493–2537
- Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 90–94). Association 15 for Computational Linguistics.
  Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012). Probase: A probabilistic taxonomy for text understanding. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 481–492). ACM.
- Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 42–49). ACM.
- Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 233–242). ACM.
- Gupta, D., Lenka, P., Bedi, H., Ekbal, A., & Bhattacharyya, P. (2017). Iitp at ijcnlp-2017 task 4: Auto analysis of customer feedback using cnn and gru network. In Proceedings of the IJCNLP 2017, Shared Tasks (pp. 184–193).
- Plank, B. (2017). All-in-1: Short text classification with one model for all languages. In Proceedings of the International Joint Conference on Natural Language Processing (Shared Task 4). Taipei, Taiwan: Association for Computational Linguistics. Post, M., & Bergsma, S. (2013).

RESEARCH PLAN



AtlantTIC