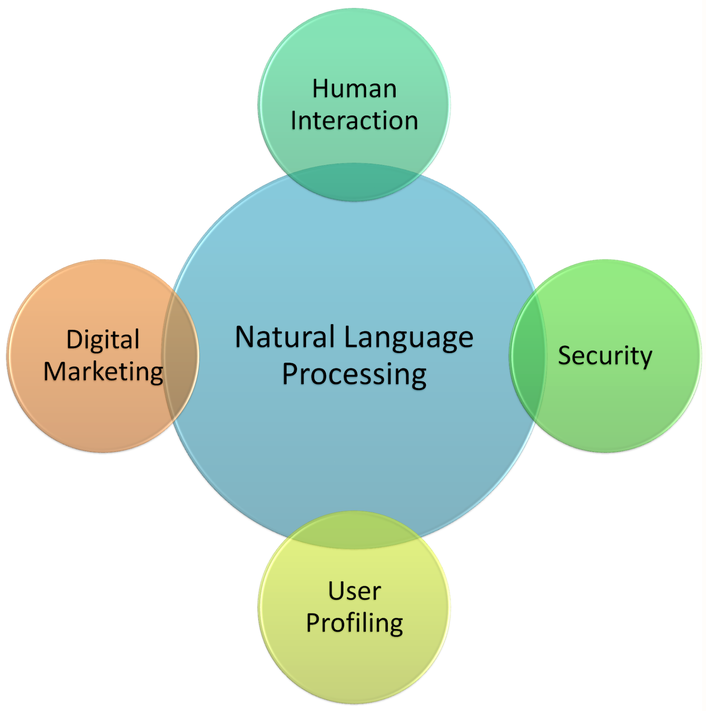
Contribution to research new models of knowledge extraction on BigData systems

Héctor Cerezo-Costas, Advisor: F.Javier González-Castaño[1]

[1]Department of Telematics Engineering, University of Vigo

Motivation

Natural Language Processing (NLP) has a wide range of applications such as:



Human performance exceed computers in many complex NLP tasks:

Nonetheless computers are faster and they are able to solve problems at web-scale.

Thesis Objectives

- Objective 1: Research in new unsupervised algorithms for the application in NLP tasks.
- Objective 2: Development of bigdata algorithms to solve NLP problems in the Terabyte-scale.
- Objective 3: Research in new technologies for fast adaptation in different context of text mining models.

Research Plan (Next Year)

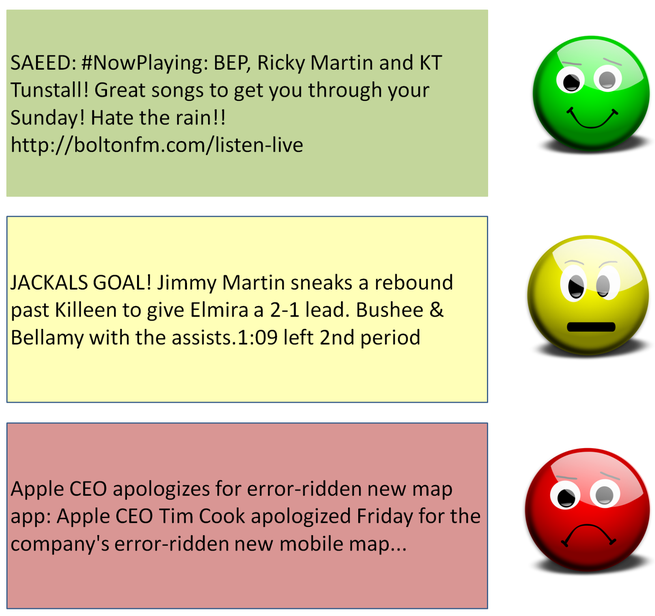
Ongoing Work

Participation in a Sentiment Analysis Competition (SemEval 2015)

We have taken part in the following competition SemEval-2015 Task 10 Subtask B: Sentiment Analysis in Twitter [1].

Goal:

- Given a message from Twitter classify it as positive, negative or neutral.



General Approach

- Supervised Strategy with Logistic Regression
- Ensemble of classifiers with majority voting strategy
- CRFs for complex feature extraction: negation, comparison, adversative clauses, etc [2, 3]

The steps performed by the system are:

1. Preprocessing Step: emoticon substitution, multiword hashtag splittage, mentions and URL substitutions, etc
2. Data Tagging: polarity dictionaries, verb reversal detection, etc.
3. PoS data extraction
4. Syntactic Information Extraction: detection of negation, adversative or polarity reversal scopes using CRFs
5. Feature extraction and classification of sentences

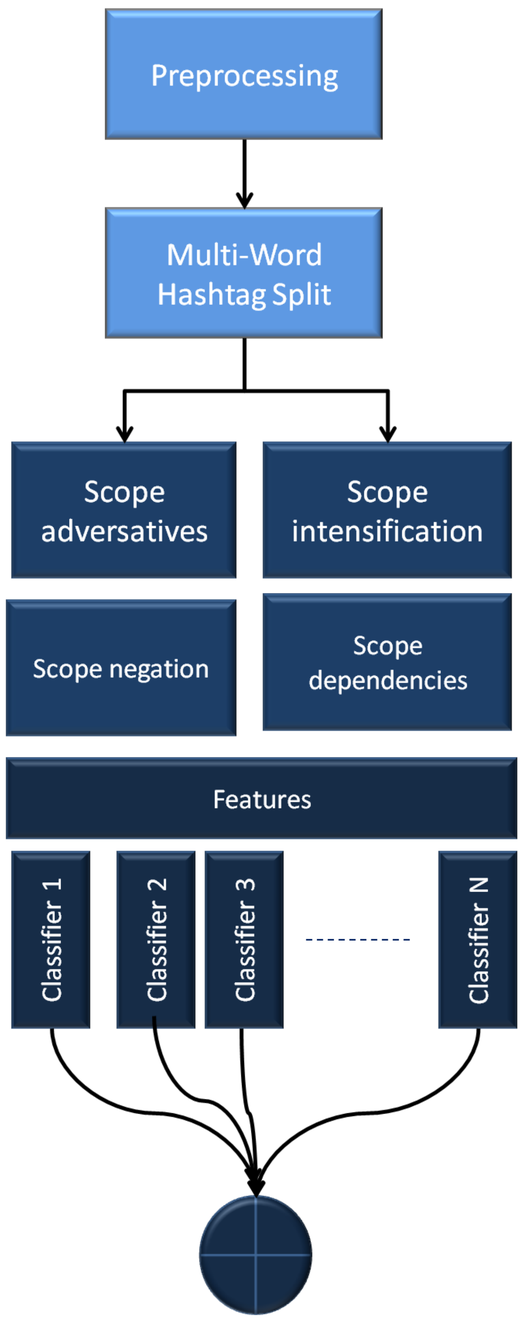


Figure 1 : Architecture of the system.

Results

Test	F-score
LiveJournal 2014	72.63
SMS 2013	61.97
Twitter 2013	65.29
Twitter 2014	66.87
Twitter 2014 sarcasm	59.11
Twitter 2015	60.62
Twitter 2015 sarcasm	56.45

Table 1 : Performance in progress and input test.

- 16th position out of 40 competitors in both sarcasm and regular 2015 datasets.
- 1st position in 2014 Tweet Sarcasm dataset.
- Generalized degradation between 2014 and 2015 performance results.

References

[1] S. Rosenthal, P. Nakov, S. Kiritchenko, S.M. Mohammad, A. Ritter, and V. Stoyanov. 2015. Semeval-2015 Task 10: Sentiment Analysis in Twitter. *In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '2015, Denver, Colorado, June*

[2] J. Lafferty, A. McCallum, and F. CN Pereira. 2001. Conditional random fields: Probabilistic models for Segmenting and Labeling Sequence Data

[3] E. Lapponi, E. Velldal, L. Øvrelid, and J. Read. 2012b. Uio 2: Sequence-Labeling Negation Using Dependency Features. *In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1.pages 319–327.*

Workshop on Monitoring PhD Student Progress, 16 June 2015, Vigo, Spain