MONDRAGON UNIBERTSITATEA
GOI ESKOLA POLITEKNIKOA
FACULTY OF ENGINEERING

Novel Comment Spam Filtering Method on Youtube: Sentiment Analysis and Personality Recognition

Rome, June 2017

Enaitz Ezpeleta, Iñaki Garitano, Ignacio Arenaza-Nuño, José María Gómez Hidalgo, and Urko Zurutuza

Electronics and Computing Department
Faculty of Engineering – Mondragon University
@mu_gep

Outline

1. Introduction

2. Proposed method

3. Sentiment Analysis

4. Personality Recognition

5. Combination of Sentiment Analysis and Personality Recognition

6. Conclusions

Introduction

Motivation

Online Social Networks popularity

- Facebook reached 1.65 billion monthly active users as of March 31, 2016 [1].
- Youtube has counted over a billion users in 2016 [2].
- Twitter has 310 million monthly active users as of March 31, 2016 [3].

[1] http://newsroom.fb.com/company-info/

[2] https://www.youtube.com/yt/press/statistics.html

[3] https://about.twitter.com/company

Motivation

Spam

- Unsolicited email campaigns remain as one of the biggest threats affecting millions of users per day.
- Spam in email traffic in Q1 2016: **56.92%** [4].
- Increase of spam in Online Social Networks.

[4] https://securelist.com/analysis/quarterly-spam-reports/74682/spam-and-phishing-in-q1-

Objective

To demonstrate that sentiment analysis and personality recognition techniques help to improve current social media spam filtering results.

Backgroud

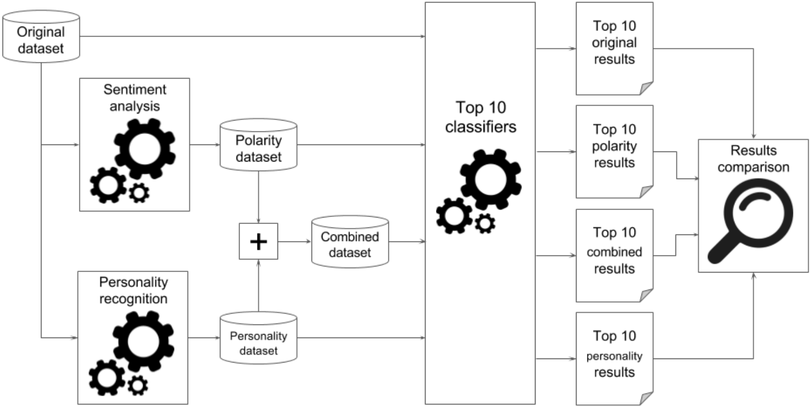
Previously published studies by Ezpeleta et. al.:

5 **Does sentiment analysis help in bayesian spam filtering?** In: Hybrid Artificial Intelligent Systems: 11th International Conference, HAIS 2016, Sevilla, Spain, April 18-20, 2016, Springer (2016)

6 **Using personality recognition techniques to improve bayesian spam filtering.** Journal Procesamiento del Lenguaje Natural (57) (2016)

Proposed method

Proposed method



PROPOSED METHOD

- All experiments are tested using 10-fold cross-validation technique.
- Results are analyzed in terms of the number of the false positives and the accuracy.
- Being the accuracy:

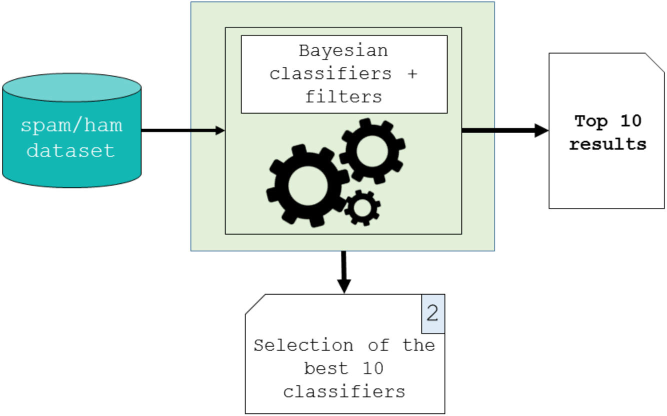
$$Accuracy = \frac{(True\ Positives + True\ Negatives)}{(Positives + Negatives)}$$

Proposed method

Youtube Comments Dataset

- Tagged dataset (ham/spam).
- 5,950,137 legitimate comments and 481,334 spam comments.
- In order to use similar number of texts to the experiments presented in [5] and [6]:
    - Subset: 1,000 spam and 3,000 ham comments
    - Randomly selected comments in English.

Social Media Spam filtering


spam/ham dataset
Bayesian classifiers + filters
Top 10 results
2
Selection of the best 10 classifiers

- Objective: To identify the best 10 spam classifiers and the best settings.

Social Media Spam filtering

#	Spam classifier	FP	Accuracy (Acc)
1	NBM.c.stwv.go.ngtok	89	82.50
2	NBMU.c.stwv.go.ngtok	89	82.50
3	NBM.stwv.go.ngtok	71	82.48
4	NBMU.stwv.go.ngtok	71	82.48
5	NBM.c.stwv.go.ngtok.stemmer	81	82.45
6	NBMU.c.stwv.go.ngtok.stemmer	81	82.45
7	NBM.stwv.go.ngtok.stemmer	64	82.35
8	NBMU.stwv.go.ngtok.stemmer	64	82.35
9	CNB.stwv.go.ngtok	125	82.30
10	CNB.stwv.go.ngtok.stemmer	109	82.28

Table 1: Results of the best ten classifiers

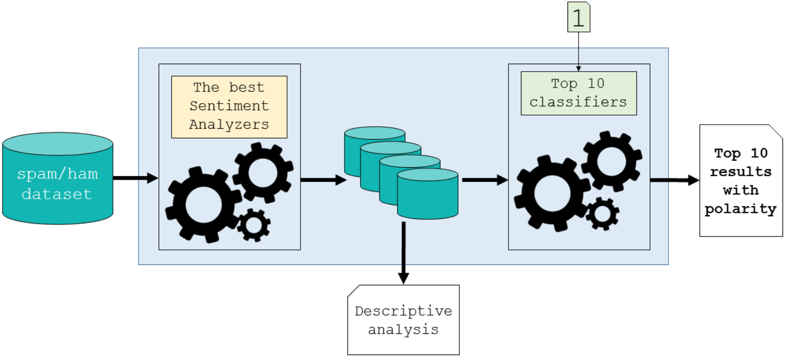
Sentiment Analysis

Definition

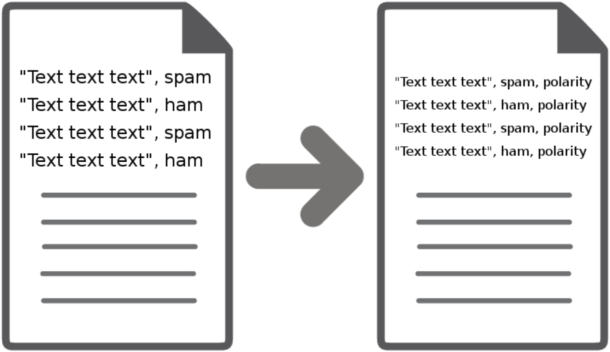
Sentiment Analysis

- ”The process of computationally identifying and categorizing opinions expressed in a piece of text.” *[Oxford Dictionaries]*
- Useful to classify the polarity of a given text (positive, negative, neutral).

Sentiment Analysis



Sentiment Analysis

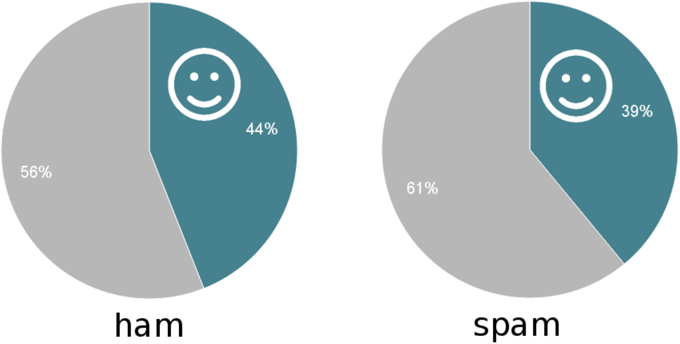


"Text text text", spam
"Text text text", ham
"Text text text", spam
"Text text text", ham

"Text text text", spam, polarity
"Text text text", ham, polarity
"Text text text", spam, polarity
"Text text text", ham, polarity

Experimental Results: Descriptive experiment

- Average of the best sentiment analyzers:



Experimental Results: Predictive experiments

Youtube comments:

- Best accuracy: from 82.50% to **82.53%**.
- The accuracy is improved in half of the cases.
- The number of false positive is reduced in all cases.
- Detailed results in the paper.

Personality Recognition

Definition

Personality Recognition

”It is a psychological construct aimed at explaining the wide variety of human behaviors in terms of a few, stable and measurable characteristics.” [7]

[7] A. Vinciarelli and G. Mohammadi. A survey of personality computing. Affective Computing, IEEE Transactions on, 5(3):273–291, 2014.

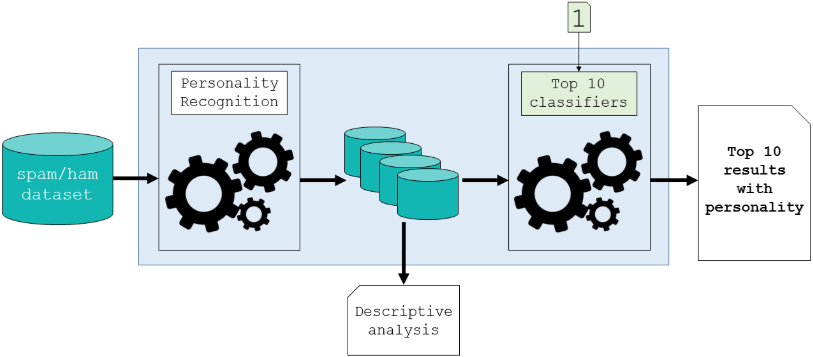
PERSONALITY RECOGNITION

Myers-Briggs personality model

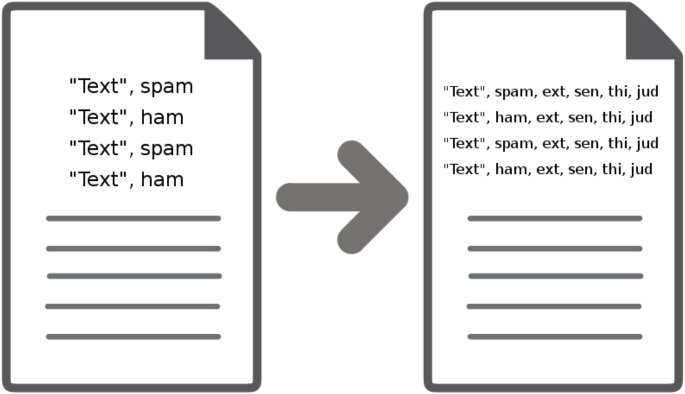
- 4 dimensions:
    - **Attitude**: Extroversion or Introversion
    - **Judging Function**: Thinking or Feeling
    - **Lifestyle**: Judging or Perceiving
    - **Perceiving Function**: Sensing or iNtuition

Publicly available web services used: `www.uClassify.com`

Personality Recognition



Personality Recognition

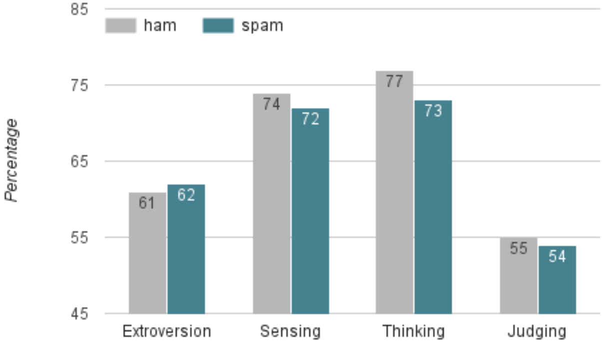


"Text", spam
"Text", ham
"Text", spam
"Text", ham

"Text", spam, ext, sen, thi, jud
"Text", ham, ext, sen, thi, jud
"Text", spam, ext, sen, thi, jud
"Text", ham, ext, sen, thi, jud

Experimental Results: Descriptive Experiment

- Personality Recognition:



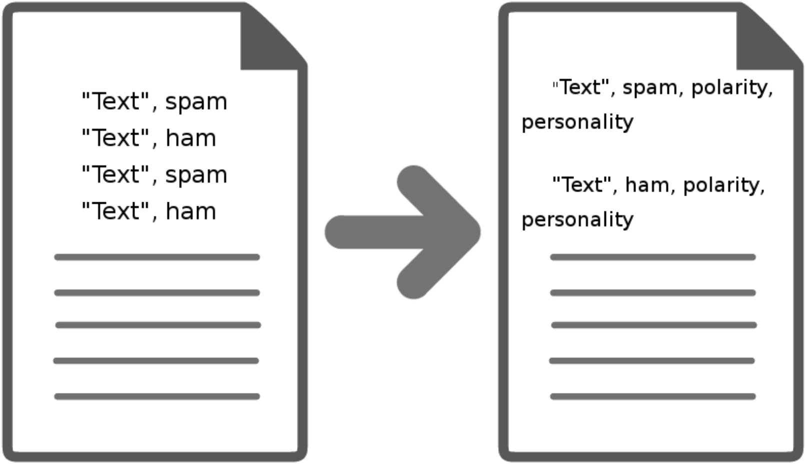
Experimental Results: Predictive experiments

Youtube comments

- Using all personality dimensions:
  - Accuracy is improved in two cases.
  - Huge reduction of the number of false positive.
- Using only the dimension Thinking:
  - Accuracies: 4 improved, 1 equalized and 5 worsened.
  - The number of false positives is reduced in all cases.

Combination of Sentiment Analysis and Personality Recognition

Combination

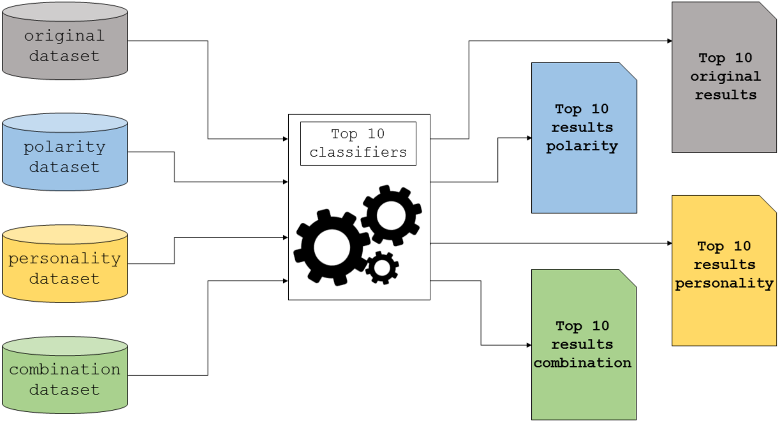


"Text", spam
"Text", ham
"Text", spam
"Text", ham

"Text", spam, polarity, personality

"Text", ham, polarity, personality

Combination

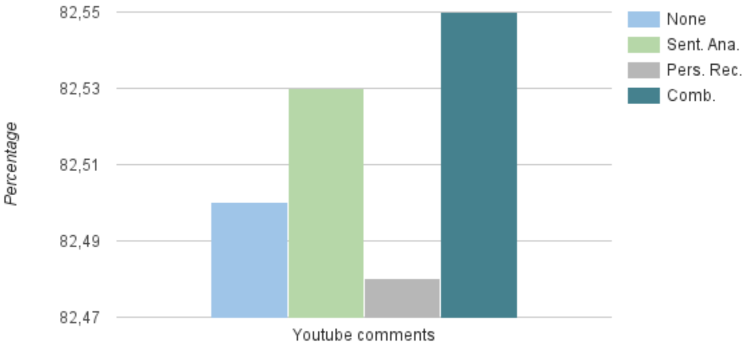


Experimental Results: Social Media Spam

	Used technique								
	None		Polarity		Personality		Comb		FP red. (%)
#	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	
1	89	**82.50**	83	82.30	76	82.38	**71**	82.30	20.22
2	89	**82.50**	83	82.30	70	82.43	**66**	82.30	25.84
3	71	**82.48**	67	82.33	61	82.35	**57**	82.20	19.72
4	71	**82.48**	67	82.33	56	82.35	**51**	82.23	28.17
5	81	82.45	74	**82.53**	69	82.48	**60**	82.48	25.93
6	81	82.45	74	82.53	65	82.48	**53**	**82.55**	34.57
7	64	82.35	59	82.20	56	**82.40**	**51**	82.18	20.31
8	64	**82.35**	59	82.20	52	82.28	**46**	82.13	28.13
9	125	82.30	104	82.40	100	82.30	**84**	**82.50**	32.80
10	109	82.28	94	82.35	87	**82.45**	**75**	82.43	31.19

Experimental Results: Social Media Spam: Summary

- The best accuracy:



- Reduction of the number of false positives in all cases.

Conclusions

Conclusions

1. We have demonstrated that sentiment analysis and personality recognition of the texts can help to detect spam in Online Social Networks.
2. In most of the cases the results are improved in both terms: accuracy and the number of the false positives.
3. Despite the difference in the accuracy percentage does not seem to be relevant, if we take into account the amount of real social spam traffic, the improvement is significant.
4. This work demonstrates that the more information about the content of the texts is added to the dataset, the better results are obtained.

Novel Comment Spam Filtering Method on Youtube: Sentiment Analysis and Personality Recognition

-

Enaitz Ezpeleta, Iñaki Garitano, Ignacio Arenaza-Nuño, José María Gómez Hidalgo, and Urko Zurutuza