International Journal of Uncertainty,
Fuzziness and Knowledge-Based Systems
Vol. 25, Suppl. 2 (December 2017) 175–189

DOI: 10.1142/S0218488517400177

World Scientific
www.worldscientific.com

Short Messages Spam Filtering Combining Personality Recognition and Sentiment Analysis

Enaitz Ezpeleta*, Iñaki Garitano† and Urko Zurutuza‡
Electronics and Computing Department, Mondragon University,
Goiru 2, 20500 Arrasate-Mondragón, Spain
**eezpeleta@mondragon.edu*
†*igaritano@mondragon.edu*
‡*uzurutuza@mondragon.edu*

José María Gómez Hidalgo
Pragsis Technologies, Manuel Tovar,
43-53, Fuencarral 28034, Madrid, Spain
jmgomez@pragsis.com

Received 17 February 2017
Revised 26 September 2017

Currently, short communication channels are growing up due to the huge increase in the number of smartphones and online social networks users. This growth attracts malicious campaigns, such as spam campaigns, that are a direct threat to the security and privacy of the users. While most researches are focused on automatic text classification, in this work we demonstrate the possibility of improving current short messages spam detection systems using a novel method. We combine personality recognition and sentiment analysis techniques to analyze Short Message Services (SMS) texts. We enrich a publicly available dataset adding these features, first separately and after in combination, of each message to the dataset, creating new datasets. We apply several combinations of the best SMS spam classifiers and filters to each dataset in order to compare the results of each one. Taking into account the experimental results we analyze the real influence of each feature and the combination of both. At the end, the best results are improved in terms of accuracy, reaching to a 99.01% and the number of false positive is reduced.

Keywords: SPAM; polarity; personality; SMS; sentiment analysis; security.

1. Introduction

In the same way that smartphones and online social networks (OSN) are growing up, short messages traffic is increasing all over the world. For example, 6.1 billion people used an SMS-capable mobile phone on June 2015, what means that SMS

messages can reach more than 6 billion users.[a] In the same way WhatsApp, one of the most famous instant messaging application, reached 1 billion users in 2016.[b]

This growth became these communication methods in a very attractive objective to malicious organizations, and more and more illegal activities are being carried out through those systems.[c] For example, Spanish police busted a gang that made at least 5 million Euros over last decade from a premium-rate SMS messaging scam.[d]

Malicious campaigns in SMS communication systems are specially effective due to the phenomenal opening rate of 98% (for instance, email marketing reports a 22% open rate).[e] This demonstrates that there are billions of users whose privacy can be threaten sending an unsolicited instant short message (For example: SMS, WhatsApp message...). Currently, with the 20-30% of all SMS traffic being sent in China and India, SMS spam is an emerging problem, specially in Asia.[f]

During the last years, several tools and systems have been proposed by researchers to deal with this problem. Most research initiatives are focused on automatic text classification, but no one take sentiment analysis or personality recognition techniques into account.

The main objective of this paper is to analyze the influence of these techniques in short instant messages spam filtering. It also aims to provide means to prove that the combination of polarity and personality dimensions can improve the results obtained previously.[1,2] In these studies the mentioned techniques were applied individually. Using the same datasets used in these papers, we focus on SMS messages, which are structurally similar to other instant short messages.

The remainder of this paper is organized as follows. Section 2 describes the previous work conducted in the area of short messages spam filtering techniques, in personality recognition and in sentiment analysis. Section 3 describes the process of the aforementioned experiments, regarding Bayesian short messages spam filtering and short messages spam filtering using the personality dimensions and the polarity. In Section 4, the obtained results are described, comparing Bayesian filtering results and the filtering results using the personality and the polarity features. Finally, we summarize our findings and give conclusions in Section 5.

2. Related Work

Several studies related to these topics have been published during the last years. In this Section we survey previous SMS spam, personality recognition and sentiment analysis contributions.

[a]https://www.yumpu.com/en/document/view/54468786/sms-the-language-of-6-billion-people
[b]https://blog.whatsapp.com/616/One-billion/
[c]https://www.clxcommunications.com/blog/2016/12/
nine-statistics-outline-problem-fraud-mobile-messaging-industry/
[d]http://goo.gl/WsNBVb
[e]http://goo.gl/CaxweY
[f]https://www.yumpu.com/en/document/view/22110268/
sms-spam-and-mobile-messaging-attacks-introduction-gsma/2

2.1. *SMS spam*

During the last years malicious users have detected that instant message services are suitable platforms to perform malicious activities, specially attracted by the huge amount of users these cope with. In this work we are focusing specially on SMS messages. Those are structurally similar to other currently more consumed short message applications such as WhatsApp, Line or even Twitter. Our decision to focus on certain messages is principally based on the public access to labelled datasets needed generate and validate classification models. This provides the possibility of comparing our results with previous works. We also base our decision on the fact that SMS spam is a real and emerging problem in countries of big population,[g] and also used by people of countries where SMS services are not charged by mobile operators.

Delany *et al.*[3] presented a survey on filtering SMS spam and showed recent developments in SMS spam filtering. Also a brief discussion about publicly available corpus and availability for future research in the area are shown.

Almeida *et al.*[4] compare different machine learning methods and indicated that Support Vector Machine technique was the best one during their study. They obtained an accuracy of 97.64% using this method. Furthermore, they offer a public and non-encoded SMS spam collection that can be used by the community. This study brings us the possibility to test with the same dataset and to compare results.

In other recent studies two-level classifiers are used to obtain better results in classifying spam.[5,6] In this study we are going to focus on improving one-level learning-based classifiers.

2.2. *Personality recognition*

Personality is a psychological construct aimed at explaining the wide variety of human behaviors in terms of a few, stable and measurable individual characteristics.[7] Celli and Poesio[8] explain two main models to formalize personality have been defined: Myers-Briggs personality model,[9] which defines the personality using four dimensions: Extroversion or Introversion, Thinking or Felling, Judging or Perceiving and Sensing or iNtuition; and the Big Five model[10] which divides the personality in 5 traits: Openness to experience, Conscientiousness, Extroversion, Agreeableness and Neuroticism.

Every text contains a lot of information about the personality of the authors, being this the reason that personality recognition became a potential tool for Natural Language Processing.[11] During the last years, different studies in personality recognition in blogs,[12] offline texts[11] or OSNs[13,14] have been published.

Shen *et al.*[15] prove that personality prediction is feasible, and their email feature set can predict personality with reasonable accuracies. This work shows that it is possible to predict the personality of a writer using email messages.

[g]https://www.yumpu.com/en/document/view/22110268/sms-spam-and-mobile-messaging-attacks-introduction-gsma/2

Although these techniques have been never used in the field of SMS spam detection, researchers present the relationship between personality traits and deceptive communication.[16,17]

2.3. *Sentiment analysis*

Natural Language Processing (NLP) techniques are becoming more and more useful for spam filtering, using sender information and text content based NLP techniques.[18]

Researchers confirmed that it is possible to create an application or a system to detect spam in different formats using text mining techniques and semantic language models respectively.[19,20]

Among all NLP techniques, we focus on the use of Sentiment Analysis (SA) to improve the detection of illegitimate short instant messages. This is a different strategy if we compare with the traditional short spam detection techniques, which focus on automatic text classification, but do not take SA into account.

During the last years SA has been used in several research areas, although there has been a continued interest for a while. Liu *et al.*[21] described the most important research opportunities related to SA. Based on that, we select document sentiment classification topic as a possible option to short messages filtering.

This area aims at defining if a document is positive or negative based on its content.[22] In order to improve the classification into positive, negative or neutral, other studies propose supervised learning techniques[23] or unsupervised learning techniques based on opinion words or phrases.[24]

Different tools with the objective of helping during the sentiment classification have been proposed in the last years. Lexicon-based methods are interesting tools for our work. Those methods are used to extract the polarity of a certain word or phrase. In[25] a comparison between 8 popular sentiment analysis methods is presented and the author develops a combined method to improve the results. Centered on short messages, Musto *et al.*[26] described a comparison between lexicon-based approaches.

Taking into account those comparisons, we decided to use the publicly available dictionary called *SentiWordNet*. The last version of this tool was presented in 2010,[27] which is an improved version of the first dictionary.[28]

3. Proposed Method: Combination of Personality Recognition and Sentiment Analysis

Taking as a baseline the previously presented studies,[1,2] the objective of this work is to combine the two techniques used in these papers in order to improve the spam filtering results.

To do that, having an original dataset: (1) we apply personality recognition technique to create a second dataset with this feature; (2) we apply sentiment analysis classifiers to the original dataset and we add the obtained polarity, in

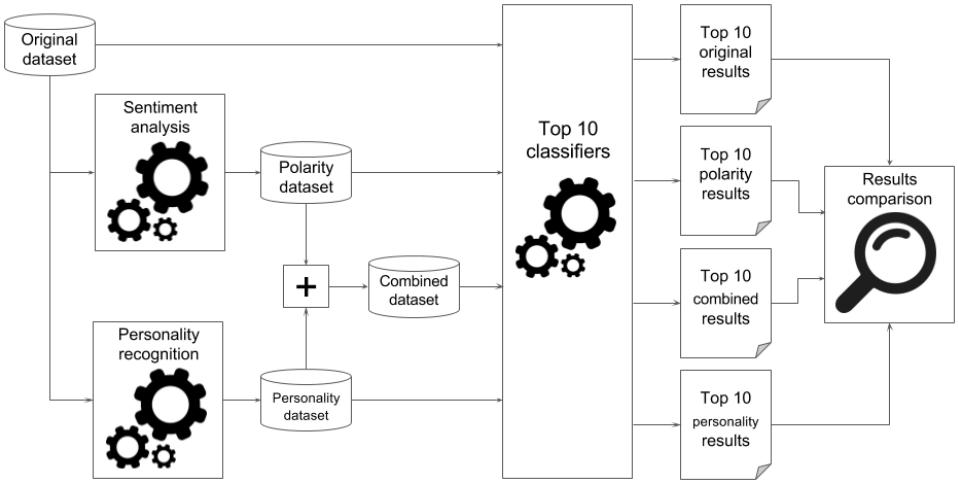

Fig. 1. Novel SMS Spam filtering method.

order to create a third dataset; (3) we combine both techniques in the original messages and we create a combined fourth dataset; (4) having these four different datasets, we apply the best ten spam filtering classifiers identified in Ref. 1 to each dataset; (5) finally, the top results of each dataset are analyzed.

In the Fig. 1 the full procedure is described.

During those experiments 10-fold cross-validation technique is used, and the results are analyzed in terms of the number of false positives and the accuracy. Accuracy is the percentage of testing set examples correctly classified by the classifier.

3.1. *Datasets*

In this work two publicly available dataset are used:

- *SMS Spam Collection v.1*[h] (called *SMSSpam* in this paper):[4] It is composed of 5,574 English, real and non-enconded messages, tagged as being legitimate (ham) or spam. Specifically, it contains 747 spam messages and 4,827 ham messages. This dataset is used to carry out the two spam filtering experiments.
- *British English SMS corpora*[i] (called *BritishSMS* in this paper):[29] This dataset contains 875 SMS messages labelled in terms of spam. There are 450 legitimate SMS messages, and 425 spam SMS messages in this dataset. During this study, we use this dataset to validate the results of the previous dataset, repeating the experiments workflow.

[h]http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/
[i]https://goo.gl/UUgl4X

3.2. *SMS spam filtering*

To analyze if the combination of different techniques improve Bayesian spam filtering and the individual application of each technique, baseline results are needed.

The best ten classifiers for spam filtering are identified taking into account the results obtained in Ref. 1. These results are shown in the Table 1.

Table 1. Top10 Bayesian classifiers.

#	Spam classifier	TP	TN	FP	FN	Acc
1	NBMU.i.c.stwv.go.ngtok	711	4,799	28	36	98.85
2	NBMU.i.t.c.stwv.go.ngtok	708	4800	27	39	98.81
3	NBM.i.t.c.stwv.go.ngtok	711	4,795	32	36	98.78
4	NBMU.i.t.c.stwv.go.ngtok.stemmer	702	4,804	23	45	98.78
5	NBM.c.stwv.go.wtok	691	4,814	13	56	98.76
6	NBM.i.t.c.stwv.go.ngtok.stemmer	712	4,793	34	35	98.76
7	NBMU.c.stwv.go.wtok	691	4,814	13	56	98.76
8	CNB.i.t.c.stwv.go.ngtok.stemmer	713	4,790	37	34	98.73
9	NBM.i.c.stwv.go.ngtok	713	4,790	37	34	98.73
10	NBM.i.c.stwv.go.ngtok.stemmer	712	4,791	36	35	98.73

During this paper, our main objective is to improve these results using the selected classifiers. To understand the settings of each classifier, Table 2 shows the nomenclatures used.

Table 2. Nomenclatures.

	Meaning		Meaning
CNB	Complement Naive Bayes	.stwv	String to Word Vector
NBM	Naive Bayes Multinomial	.go	General options
NBMU	Naive Bayes Multinomial Updatable	.wtok	Word Tokenizer
.c	idft F, tft F, outwc T	.ngtok	NGram Tokenizer 1-3
.i.c	idft T, tft F, outwc T	.stemmer	Stemmer
.i.t.c	idft T, tft T, outwc T	.igain	Attribute selection using InfoGainAttributeEval

3.3. *SMS spam filtering using personality recognition*

Following the procedure presented in Ref. 1, we use one of the most trusted personality model: Myers-Briggs personality model. This model is composed of four different dimensions (Extroversion or Introversion, Thinking or Feeling, Judging or Perceiving and Sensing or iNtuition), which are mandatory in order to determine the personality. To calculate the dimensions of each text, we use publicly available machine learning web services for text classification hosted in *uClassify*.[j] Among

[j]https://www.uclassify.com

all the possibilities offered in this website, we focus on the Myers-Briggs functions developed by Mattias Östmar.

As the author explains, each function determines a certain dimension of the personality type according to Myers-Briggs personality model. The analysis is based on the writing style and should not be confused with the Myers-Briggs Type Indicator (MBTI) which determines personality type based on self-assessment questionnaires. Training texts are manually selected based on personality and writing style according to Jensen.[30] Those are the used functions:

- *Myers-Briggs Attitude:* Analyzes the Extroversion or Introversion dimension.
- *Myers-Briggs Judging Function:* Determines the Thinking or Feeling dimension.
- *Myers-Briggs Lifestyle:* Determines the Judging or Perceiving dimension.
- *Myers-Briggs Perceiving Function:* Determines the Sensing or iNtuition dimension.

Each function returns a float within the range [0.0, 1.0] per each pair of characteristics of the dimension. For example, if we test a certain text and we obtain X value for Extroversion, the value for Introversion is 1-X. Thus, we only record one value per each function: Extroversion, Sensing, Thinking and Judging.

Those four values of each SMS message are added to the original dataset in order to create a new dataset. During the experiments, this new dataset is used in order to see the influence of the personality dimensions during the SMS spam filtering. To do that, we apply the top ten classifiers mentioned previously to the original dataset and to the new one, and we compare the results.

3.4. *SMS spam filtering using sentiment analysis*

The main objective of this part is to add the polarity of each message to the original dataset. To do that, we use the procedure and configuration options shown in Ref. 2 where the best sentiment classifiers were identified to carry out the experiments. Based on the accuracies presented in the mentioned paper, the best three classifiers are selected (*TextBlob 0.05*, *TextBlob 0.1* and *TextBlob -0.05*) in order to use those ones to annotate the messages included in *SMS Spam Collection v.1* which has not been annotated for sentiment. As a result, we obtain three new datasets (one per each classifier). The original one and the new three ones are used in the next experiments.

3.5. *SMS spam filtering combining personality recognition and sentiment analysis*

Being our objective to explore the possibilities to improve the previously published results,[1,2] where authors used personality dimensions and polarity feature respectively in Bayesian spam filtering. We combine both techniques, we create a new dataset adding the personality dimensions and the polarity, obtained applying the

best sentiment classifier, of each SMS message to the original dataset. Finally, we apply the best ten spam filtering classifiers to compare all the results.

4. Experimental Results

In this section, the results obtained during the previously explained experiments are shown. To carry out these experiments the dataset called *SMS Spam Collection v.1* is used, and to validate the results, the other one: *BritishSMS*.

4.1. *Descriptive analysis*

Once the dataset is selected, we perform a descriptive experiment of the dataset. The objective of this step is on the one hand to perform an analysis to extract the personality dimensions from each SMS messages, and on the other hand to analyze the polarity of the messages.

The personality dimensions of each messages is extracted applying the previously explained personality recognition technique. In this point a new dataset is created by inserting the personality features extracted during the analysis. Finally the statistics about the personality dimensions in SMS messages are calculated. Those statistics are presented in Fig. 2.

Results show that all the dimensions of the personality model have a different distribution depending on the text type. At this point we can confirm that the way SMS messages are written (spam/ham) varies. Furthermore, from the perspective of the effect of personality on deceptive communication the interesting thing is the difference in spam/ham messages with respect to the judging personality trait.[16]

To analyze the polarity of the messages, the previously selected sentiment classifiers are used. Like in the personality part, the polarity extracted during the analysis is inserted in the dataset, creating three new datasets (one per each classifier).

Figure 2 results show that spam messages are mostly positive while ham messages are more negative. This means that there is a difference between spam and ham messages in terms of polarity, so it can be helpful for improving SMS spam filtering.

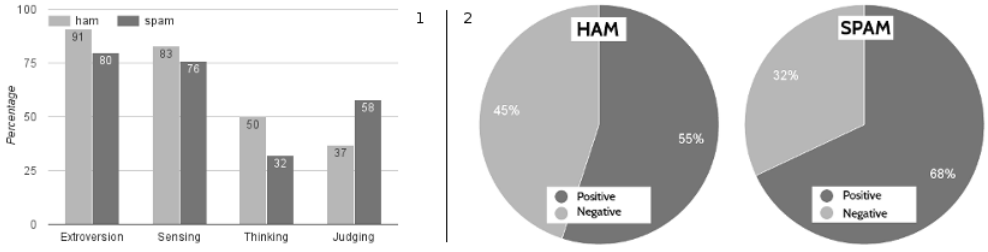

Fig. 2. (1) Analysis of the personality dimensions in SMS messages. (2) Polarity comparison between spam and ham SMS messages.

4.2. *Spam filtering: baseline results*

The results obtained applying the best 10 filtering classifiers to the previously created datasets (original, 3 with polarity, with personality) are shown in Table 3.

Table 3. Comparing original results with the results obtained using sentiment analyzers and personality recognition techniques.

	Sentiment analyzer									
	None		*Tb 005*		*Tb 01*		*Tb -005*		*Personality*	
#	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**
1	28	98.85	36	98.73	36	98.73	35	98.74	26	98.83
2	27	98.82	17	98.60	16	98.71	8	98.76	19	**98.94**
3	32	98.78	37	98.74	37	98.74	33	98.78	28	98.85
4	23	98.78	36	98.71	36	98.71	34	98.74	19	98.80
5	13	98.76	33	98.78	32	98.80	28	98.85	5	98.78
6	34	98.76	34	98.74	33	98.74	32	98.76	30	98.78
7	13	98.76	17	98.60	16	98.71	8	98.76	**3**	98.49
8	37	98.73	28	98.85	28	98.85	27	98.82	34	98.76
9	37	98.73	26	98.85	25	98.87	22	**98.91**	33	98.78
10	36	98.73	23	98.80	22	98.82	19	98.82	33	98.76

Table 3 shows that in half of the cases, polarity helps to improve the accuracy. The application of the Bayesian Logistic Regression classifier to the dataset created by *TextBlob-005* (sentiment analyzer) improves the best result. The use of polarity-driven features improve the accuracy from 98.85 % to 98.91%.

Furthermore, in some cases where better accuracy is not obtained, polarity helps to reduce the number of false positives. For instance, in the two cases where a percentage of 98.76% is obtained, the number of false positives is reduced from 27 to 8 in one case, and from 13 to 8 in the other.

Analyzing the information shown in Table 3 we see that personality feature also improves almost all the original results. In terms of accuracy a 98.94% is reached improving the best result obtained applying the classifiers to the original dataset. Only in two cases the accuracy is worst using personality score than without personality, but the false positive number is reduced in both (from 27 to 19 and from 13 to 3). However, those are not the unique cases where the number of false positives is reduced, because using the personality feature in all cases of Table 3 this number is improved.

4.3. *Second dataset: validation*

To accomplish the validation of the experiment, we select the same ten classifiers which provide the best results with the *SMSSpam* dataset, and we apply them to the *BritishSMS* dataset with and without personality feature using the 10-fold cross-validation technique. The obtained results are presented in Table 4.

Table 4. Results of the best 10 classifiers using the BritishSMS dataset.

#	Spam classifier	TP	TN	FP	FN	Acc
1	NBM.i.c.stwv.go.ngtok	408	445	5	17	97.49
2	NBM.i.t.c.stwv.go.ngtok	407	445	5	18	97.37
3	NBMU.i.c.stwv.go.ngtok	409	443	7	16	97.37
4	NBMU.i.t.c.stwv.go.ngtok	408	444	6	17	97.37
5	NBMU.i.t.c.stwv.go.ngtok.stemmer	409	441	9	16	97.14
6	NBM.i.c.stwv.go.ngtok.stemmer	407	442	8	18	97.03
7	CNB.i.t.c.stwv.go.ngtok.stemmer	406	442	8	19	96.91
8	NBM.i.t.c.stwv.go.ngtok.stemmer	406	442	8	19	96.91
9	NBM.c.stwv.go.wtok	402	441	9	23	96.34
10	NBMU.c.stwv.go.wtok	402	441	9	23	96.34

In the next step, we carry out a sentiment analysis of the *BritishSMS* dataset, using the same three sentiment analyzers used in the previous dataset, and we add the polarity feature to the original dataset. Doing that, three new tagged dataset are created. The same process is also followed in term of personality adding this feature to the original dataset.

As in the previous experiment, the same ten classifiers are applied to the new datasets in order to compare the results with the results presented in Table 4.

Analyzing Table 5, we can see that although the top result is not improved in terms of accuracy, we reach the same accuracy in different cases. And almost in all the cases the results are better or the same using the polarity feature. Also, analyzing the number of false positives, it is possible to see that the results are better or at least the same in all cases. Taking into account that the dataset is relatively small (875 SMSs), any improvement in percentages or in numbers is significant.

Analyzing the personality column, we can conclude that the use of personality features improve the spam classification, thus validating the hypothesis. Although

Table 5. Comparing original results with the results obtained using sentiment analyzers and personality recognition techniques. Second dataset.

	Sentiment analyzer									
	None		*Tb 005*		*Tb 01*		*Tb -005*		*Personality*	
#	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**
1	5	97.49	5	97.49	5	97.49	5	97.49	5	97.49
2	5	97.37	5	97.37	5	97.37	5	97.37	3	97.49
3	7	97.37	7	97.37	7	97.37	6	97.49	6	97.49
4	6	97.37	6	97.37	6	97.37	6	97.37	6	97.37
5	9	97.14	9	97.14	9	97.14	9	97.03	9	97.03
6	8	97.03	8	97.03	8	97.03	8	97.03	7	97.14
7	8	96.91	8	96.91	8	96.91	7	97.03	6	97.14
8	8	96.91	8	96.91	8	96.91	7	97.03	6	97.14
9	9	96.34	9	96.34	9	96.34	6	96.57	**2**	96.80
10	9	96.34	9	96.34	9	96.34	6	96.57	9	96.46

the best result is not improved in terms of accuracy, we reach the same accuracy in three different classifiers, and in almost all the cases results are improved. In addition, if we analyze the false positives results, those are also improved in most of the cases.

4.4. *Novel method: SMS spam filtering combining personality recognition and sentiment analysis*

Aiming at analyzing the new method proposed in Section 3, two experiments are carried out using the *SMSSpam* and the *BritishSMS* dataset.

Table 6 shows the results obtained applying the best classifiers to the *SMSSpam* dataset. In this case, the sentiment analyzer *TextBlob -0.05* and all the dimensions of the personality recognition model are used to create the combined dataset.

Table 6. Comparison of the best classifiers using the dataset *SMSSpam*.

	Used technique							
	None		*TB -0.05*		*Pers*		*Comb*	
Spam classifier	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**
NBMU.i.c.stwv.go.ngtok	28	98.85	35	98.74	26	98.83	23	**98.89**
NBMU.i.t.c.stwv.go.ngtok	27	98.82	8	98.76	19	98.94	15	**99.01**
NBM.i.t.c.stwv.go.ngtok	32	98.78	33	98.78	28	98.85	26	**98.89**
NBMU.i.t.c.stwv.go. .ngtok.stemmer	23	98.78	34	98.74	19	98.80	14	**98.87**
NBM.c.stwv.go.wtok	13	98.76	28	**98.85**	5	98.78	**4**	98.74
NBM.i.t.c.stwv.go. .ngtok.stemmer	34	98.76	32	98.76	30	98.78	25	**98.85**
NBMU.c.stwv.go.wtok	13	98.76	8	**98.76**	3	98.49	**3**	98.44
CNBi.t.c.stwv.go. .ngtok.stemmer	37	98.73	27	**98.82**	34	98.76	31	98.80
NBM.i.c.stwv.go.ngtok	37	98.73	22	**98.91**	33	98.78	31	98.82
NBM.i.c.stwv.go. .ngtok.stemmer	36	98.73	19	**98.82**	33	98.76	33	98.76

Almost all the original results are improved in terms of accuracy or the number of false positives. Moreover, the best accuracy results of the original dataset (98.85%), are improved with sentiment analysis (98.91%) and personality features (98.94%); reaching to a 99.01% with the combination of both features.

In addition, to validate those results a second dataset is used and the results are shown in Table 7. Once again, the best accuracy is improved, obtaining a 97.6% of accuracy, and reducing the number of false positives in most of the classifiers.

In order to compare the best accuracies obtained during the different experiments, we summarize them in Fig. 3.

Table 7. Comparison of the best classifiers using the dataset *BritishSMS*.

	None		*TB -0.05*		*Pers*		*Comb*	
Spam classifier	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**
NBM.i.c.stwv.go.ngtok	5	97.49	5	97.49	5	97.49	5	97.49
NBM.i.t.c.stwv.go.ngtok	5	97.37	5	97.37	3	97.49	3	97.49
NBMU.i.c.stwv.go.ngtok	7	97.37	6	97.49	6	97.49	5	**97.60**
NBMU.i.t.c.stwv.go.ngtok	6	97.37	6	97.37	6	97.37	6	97.37
NBMU.i.t.c.stwv.go. .ngtok.stemmer	9	97.14	9	97.03	9	97.03	9	97.03
NBM.i.c.stwv.go. .ngtok.stemmer	8	97.03	8	97.03	7	97.14	7	97.14
CNB.i.t.c.stwv.go. .ngtok.stemmer	8	96.91	7	97.03	6	97.14	6	97.14
NBM.i.t.c.stwv.go. .ngtok.stemmer	8	96.91	7	97.03	6	97.14	6	97.14
NBM.c.stwv.go.wtok	9	96.34	6	96.57	2	96.80	**1**	96.80
NBMU.c.stwv.go.wtok	9	96.34	6	96.57	9	96.46	6	96.69

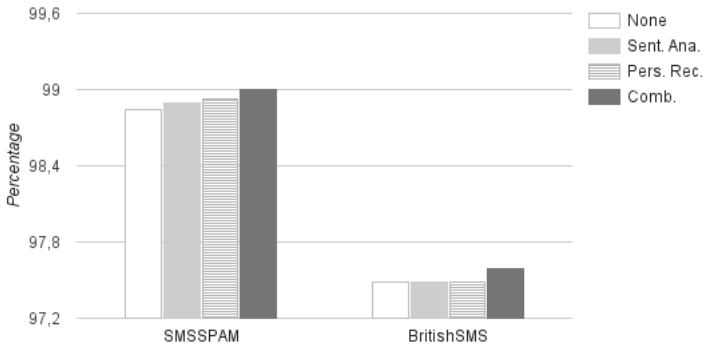

Fig. 3. Comparison of SMS spam filtering methods

5. Conclusions

This paper presents a new filtering method that gives the research community the opportunity to detect non evident intent in spam. This new method consists in using a combination of the polarity feature and the dimensions of Myers-Briggs personality model.

We added both features to the datasets, and we carried out the experiments with and without these features. With this combination we provided means to validate our hypothesis, that it is possible to identify some insights of the intention of the texts, and more spam texts are correctly classified.

As results reveal, the combination of NLP techniques help improving spam filtering in terms of accuracy and reduce the number of false positive.

Table 8. Comparison of different spam types.

	Polarity	**Personality**	**Combination**
Email[31]	BA from 99.15% to 99.21%	9/10 results improved or equalized	BA from 99.15% to 99.24%
SMS	BA from 98.85% to 98.91%	BA from 98.85% to 98.94%	BA from 98.85% to 99.01%
SMS (second dataset)	9/10 results improved or equalized	9/10 results improved or equalized	BA from 97.49% to 97.6%
Social Media[32]	– BA from 82.5% to 82.53% – Number of FP is reduced by 10% on average	– 5/10 results improved or equalized – Number of FP is reduced by 15% on average	– BA from 82.5% to 82.53% – Number of FP is reduced by 26% on average

Moreover, this method is validated in two different SMS dataset improving the best accuracy in both cases (99.01% and 97.60%) and reducing the number of false positives. Despite the difference in the percentage does not seem to be relevant, if we take into account the amount of real SMS traffic the improvement is significant. This means that sentiment analysis and personality recognition techniques are capable to highlight differences between spam and ham texts.

Moreover, as Table 8 shows, this method improved results in terms of the best accuracy (BA), best 10 results or the number of false positives (FP) applying to other types of spam. These results demonstrate that it is possible to improve spam detection results applying sentiment analysis and personality recognition techniques. Furthermore, being Online Social Networks (OSN) the new trend of research, authors applied this method to social media spam dataset. The obtained results are summarized in the Table 8. In the experiments the best accuracy is improved and the number of false positive is reduced significantly.

Acknowledgements

This work has been partially funded by the Basque Department of Education, Language policy and Culture under the project SocialSPAM (PI_2014_1_102).

We thank Mattias Östmar for the valuable tools published. And we thank Jon Kâgström (uClassify[k]) for the opportunity to use their API for research purposes.

References

1. E. Ezpeleta, U. Zurutuza and J.M.G. Hidalgo, Short messages spam filtering using personality recognition, in *Proc. 4th Spanish Conf. Information Retrieval* (*CERI '16*), New York, USA (ACM, 2016), pp. 1–7.

[k]https://www.uclassify.com

2. E. Ezpeleta, U. Zurutuza and J. M. Gómez Hidalgo, in *Short Messages Spam Filtering Using Sentiment Analysis* (Springer International Publishing, 2016), pp. 142–153.
3. S. J. Delany, M. Buckley and D. Greene, SMS spam filtering: methods and data, *Expert Systems with Applications* **39**(10) (2012) 9899–9908.
4. T. A. Almeida, J. M. Gómez Hidalgo and A. Yamakami, Contributions to the study of SMS spam filtering: new collection and results, in *Proc. 11th ACM Symposium on Document Engineering* (ACM, 2011), pp. 259–262.
5. A. Narayan and P. Saxena, The curse of 140 characters: evaluating the efficacy of SMS spam detection on android, in *Proc. Third ACM Workshop on Security and Privacy in Smartphones and Mobile Devices* (ACM, 2013), pp. 33–42.
6. N. K. Nagwani and A. Sharaff, SMS spam filtering and thread identification using bi-level text classification and clustering techniques, *Journal of Information Science* **43**(1) (2015), pp. 75–87.
7. A. Vinciarelli and G. Mohammadi, A survey of personality computing, *IEEE Transactions on Affective Computing* **5**(3) (2014) 273–291.
8. F. Celli and M. Poesio, PR2: A language independent unsupervised tool for personality recognition from text, *CoRR* **abs/1402.2796** (2014).
9. I. B. Myers and P. B. Myers, *Gifts Differing: Understanding Personality Type* (Davies-Black Publishing, 1980).
10. P. T. Costa and R. R. McCrae, Normal personality assessment in clinical practice: The NEO personality inventory, *Psychological Assessment* **4**(1) (1992) 5–13.
11. F. Mairesse, M. A. Walker, M. R. Mehl and R. K. Moore, Using linguistic cues for the automatic recognition of personality in conversation and text, *J. Artif. Int. Res.* **30**(1) (2007) 457–500.
12. J. Oberlander and S. Nowson, Whose thumb is it anyway?: Classifying author personality from weblog text, in *Proc.COLING/ACL on Main Conference Poster Sessions, COLING-ACL '06, Stroudsburg, PA, USA, Association for Computational Linguistics* (2006), pp. 627–634.
13. S. Bai, T. Zhu and L. Cheng, Big-five personality prediction based on user behaviors at social network sites, *CoRR* **abs/1204.4809** (2012).
14. F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein and W. Daelemans, *Overview of the 3rd Author Profiling Task at PAN 2015*, in Working Notes Papers of the CLEF 2015 Evaluation Labs CEUR Workshop Proceedings, CLEF and CEUR-WS.org (September 2015).
15. J. Shen, O. Brdiczka and J. Liu, Understanding email writers: Personality prediction from email messages, in *User Modeling, Adaptation, and Personalization* (Springer, 2013), pp. 318–330.
16. T. Fornaciari, F. Celli and M. Poesio, The effect of personality type on deceptive communication style, in *2013 European Intelligence and Security Informatics Conference* (*EISIC 2013*), Uppsala, Sweden, August 2013, pp. 1–6.
17. D. Hernández Fusilier, M. Montes-y Gómez, P. Rosso and R. Guzmán Cabrera, Detecting positive and negative deceptive opinions using PU-learning, *Inf. Process. Manage.* **51**(4) (2015) 433–443.
18. R. Giyanani and M. Desai, Spam detection using natural language processing, *Int. J. Computer Science Research and Technology* **1** (2013) 55–58.
19. P. F. Echeverria Briones, Z. V. Altamirano Valarezo, A. B. Pinto Astudillo and J. D. C. Sanchez Guerrero, Text mining aplicado a la clasificación y distribución automática de correo electrónico y detección de correo spam (2009).

20. R. Y. K. Lau, S. Y. Liao, R. C. W. Kwok, K. Xu, Y. Xia and Y. Li, Text mining and probabilistic language modeling for online review spam detection, *ACM Trans. Manage. Inf. Syst.* **2**(4) (2012) 1–30.
21. B. Liu and L. Zhang, A survey of opinion mining and sentiment analysis, *Mining Text Data* (2012) 415–463.
22. B. Pang and L. Lee, Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval* **2**(1-2) (2008) 1–135.
23. B. Pang, L. Lee and S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, in *Proc. ACL-02 Conference on Empirical Methods in Natural Language Processing* (*EMNLP '02*), Vol. 10, Stroudsburg, PA, USA, Association for Computational Linguistics (2002), pp. 79–86.
24. P. D. Turney, Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in *Proc. 40th Annual Meeting on Association for Computational Linguistics* (*ACL '02*), Stroudsburg, PA, USA, Association for Computational Linguistics (2002), pp. 417–424.
25. P. Gonçalves, M. Araújo, F. Benevenuto and M. Cha, Comparing and combining sentiment analysis methods, in *Proc. First ACM Conf. Online social Networks* (ACM, 2013), pp. 27–38.
26. C. Musto, G. Semeraro and M. Polignano, A comparison of lexicon-based approaches for sentiment analysis of microblog posts, *Information Filtering and Retrieval* (2014), p. 59.
27. S. Baccianella, A. Esuli and F. Sebastiani, Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in *LREC*, Vol. 10 (2010), pp. 2200–2204.
28. A. Esuli and F. Sebastiani, Sentiwordnet: A publicly available lexical resource for opinion mining, in *Proc. LREC*, Vol. 6 (Citeseer, 2006), pp. 417–422.
29. M. T. Nuruzzaman, C. Lee and D. Choi, Independent and personal SMS spam filtering, in *2011 IEEE 11th Int. Conf. Computer and Information Technology* (*CIT*), August 2011, pp. 429–435.
30. G. H. Jensen and J. K. DiTiberio, *Personality and the Teaching of Composition* (1989).
31. E. Ezpeleta, U. Zurutuza, and J. M. G. Hidalgo, Using personality recognition techniques to improve Bayesian spam filtering, *Procesamiento del Lenguaje Natural* **57** (2016) 125–132.
32. E. Ezpeleta, I. Garitano, I. Arenaza-Nuño, U. Zurutuza and J. M. G. Hidalgo, Novel comment spam filtering method on youtube: Sentiment analysis and personality recognition, in *Proc. Current Trends in Web Engineering* (*ICWE 2017*), International Workshops (Springer International Publishing, 2017).