New approaches for content-based analysis towards Online Social Network spam detection

Enaitz Ezpeleta Gallastegi

Electronics and Computing Department
Goi Eskola Politeknikoa
Mondragon Unibertsitatea

June 2016

New approaches for content-based analysis towards Online Social Network spam detection

Enaitz Ezpeleta Gallastegi

Supervisors:

Dr. Urko Zurutuza Ortega

Dr. José María Gómez Hidalgo

Thesis submitted for the degree of Doctor of Philosophy
at the University of Mondragon

Committee:
Chair: Dr. Manel Medina (Universitat Politècnica de Catalunya)
Member: Dr. Magnus Almgren (Chalmers University of Technology)
Member: Dr. José Ramón Méndez (Universidade de Vigo)
Member: Dr. Igor Santos (Universidad de Deusto)
Secretary: Dr. Iñaki Garitano (Mondragon Unibertsitatea)

June 2016

ama, aitte ta Geaxiri

Today is life, tomorrow never comes.
Matala, Crete.

Originality Statement

I declare that I am the sole author of this work. This is a true copy of the final document, including any revision which may have been ordered by my examiners. I understand that my work may be available to the public, either in the school library or in electronic format.

Abstract

Unsolicited email campaigns remain as one of the biggest threats affecting millions of users per day. Although spam filtering techniques are capable of detecting significant percentage of the spam messages, the problem is far from being solved, specially due to the total amount of spam traffic that flows over the Internet, and new potential attack vectors used by malicious users.

The deeply entrenched use of Online Social Networks (OSNs), where millions of users share unconsciously any kind of personal data, offers a very attractive channel to attackers. Those sites provide two main interesting areas for malicious activities: exploitation of the huge amount of information stored in the profiles of the users, and the possibility of targeting user addresses and user spaces through their personal profiles, groups, pages... Consequently, new type of targeted attacks are being detected in those communication means.

Being selling products, creating social alarm, creating public awareness campaigns, generating traffic with viral contents, fooling users with suspicious attachments, etc. the main purpose of spam messages, those type of communications have a specific writing style that spam filtering can take advantage of.

The main objectives of this thesis are: (i) to demonstrate that it is possible to develop new targeted attacks exploiting personalized spam campaigns using OSN information, and (ii) to design and validate novel spam detection methods that help detecting the intentionality of the messages, using natural language processing techniques, in order to classify them as spam or legitimate. Additionally, those methods must be effective also dealing with the spam that is appearing in OSNs.

To achieve the first objective a system to design and send personalized spam campaigns is proposed. We extract automatically users' public information from a well known social site. We analyze it and design different templates taking into account the preferences of the users. After that, different experiments are carried out sending typical and personalized spam. The results show that the click-through rate is considerably improved with this new strategy.

In the second part of the thesis we propose three novel spam filtering methods. Those methods aim to detect non-evident illegitimate intent in order to add valid information that is used by spam classifiers. To detect the intentionality of the texts, we hypothesize that sentiment analysis and personality recognition techniques could provide new means to differentiate spam text from legitimate one. Taking into account this assumption, we present three different methods: the first one uses sentiment analysis to extract the polarity feature of each analyzed text, thus we analyze the optimistic or pessimistic attitude of spam messages compared to legitimate texts. The second one uses personality recognition techniques to add personality dimensions (Extroversion/Introversion, Thinking/Feeling, Judging/Perceiving and Sensing/iNtuition) to the spam filtering process; and the last one is a combination of the two previously mentioned techniques.

Once the methods are described, we experimentally validate the proposed approaches in three different types of spam: email spam, SMS spam and spam from a popular OSN.

Laburpena

Hartzailearen baimenik gabe bidalitako mezuak (spam) egunean milioika erabiltzaileri eragiten dien mehatxua dira. Nahiz eta spam detekzio tresnek gero eta emaitza hobeagoak lortu, arazoa konpontzetik oso urruti dago oraindik, batez ere spam kopuruari eta erasotzaileen estrategia berriei esker.

Hori gutxi ez eta azken urteetan sare sozialek izan duten erabiltzaile gorakadaren ondorioz, non milioika erabiltzailek beraien datu pribatuak publiko egiten dituzten, gune hauek oso leku erakargarriak bilakatu dira erasotzaileentzat. Batez ere bi arlo interesgarri eskaintzen dituzte webgune hauek: profiletan pilatutako informazio guztiaren ustiapena, eta erabiltzaileekin harreman zuzena izateko erraztasuna (profil bidez, talde bidez, orrialde bidez...). Ondorioz, gero eta ekintza ilegal gehiago atzematen ari dira webgune hauetan.

Spam mezuen helburu nagusienak zerbait saldu, alarma soziala sortu, sentsibilizazio kanpainak martxan jarri, etab. izaki, mezu mota hauek eduki ohi duten idazketa mezua berauen detekziorako erabilia izan daiteke.

Lan honen helburu nagusiak ondorengoak dira: alde batetik, sare sozialetako informazio publikoa erabiliz egungo detekzio sistemak saihestuko dituen spam pertsonalizatua garatzea posible dela erakustea; eta bestetik hizkuntza naturalaren prozesamendurako teknikak erabiliz, testuen intentzionalitatea atzeman eta spam-a detektatzeko metodologia berriak garatzea. Gainera, sistema horiek sare sozialetako spam mezuekin lan egiteko gaitasuna ere izan beharko dute.

Lehen helburu hori lortzekolan honetan spam pertsonalizatua diseinatu eta bidaltzeko sistema bat aurkeztu da. Era automatikoan erabiltzaileen informazio publikoa ateratzen dugu sare sozial ospetsu batetik, ondoren informazio hori aztertu eta txantiloi ezberdinak garatzen ditugu erabiltzaileen iritziak kontuan hartuaz. Behin hori egindakoan, hainbat esperimentu burutzen ditugu spam normala eta pertsonalizatua bidaliz, bien arteko emaitzen ezberdintasuna alderatzeko.

Tesiaren bigarren zatian hiru spam atzemate metodologia berri aurkezten ditugu. Berauen helburua tribialak ez den intentzio komertziala atzeman ta hori baliatuz spam mezuak sailkatzean datza. Intentzionalitate hori lortze aldera, analisi sentimentala eta pertsonalitate detekzio teknikak erabiltzen ditugu. Modu honetan, hiru sistema ezberdin aurkezten dira hemen: lehenengoa analisi sentimentala soilik erabiliz, bigarrena lan honetarako pertsonalitate detekzio teknikek eskaintzen dutena aztertzen duena, eta azkenik, bien arteko konbinazioa.

Tresna hauek erabiliz, balidazio esperimentala burutzen da proposatutako sistemak eraginkorrak diren edo ez aztertzeko, hiru mota ezberdinetako spam-arekin lan eginez: email spam-a, SMS spam-a eta sare sozial ospetsu bateko spam-a.

Eskertzak - Acknowledgments

Badia urte batzuk bide honi ekin notzanetik... Hasiera batian gailurra oso urrin ikusi arren, pausoz pauso, kanpalekuz kanpaleku, mendi gaina zapaldu doten seinale da hitz honek idazten egotia. Baina tontorrera heltzeko bidian danak eztia malda gutxiko aldapak izaten. Horregaitxik, ibilbidian topautako oztopuak gainditzen lagundu dozten pertsonei eskerrak emoteko aukera aprobetxau nahiko nuke, eurak barik ezinezkua izango zalako gailurra lortzia, eta eurei esker holako erronketan garrantzitsuena dana lortu dotelako: **tontorrerainoko bidiaz gozatzia**.

Lehenengo ta behin, aipamen berezi bat **Urko Zurutuzai** bide hau itxeko aukeria emon, gidari lanak in eta emondako laguntza danagaitxik, baina batez be ezkortasun momentutan zure baikortasun mugagabe horrekin aurrera jarraitzeko bultzadia emotearren. En la misma linea, también me gustaría agradecer a mi codirector **José María Gómez Hidalgo**, que desde la distancia ha dedicado parte de su tiempo en que el buen fin de esta tesis sea posible. Bestalde, eskerrak emon baitxa **Mondragon Unibertsitatiai** lan hau burutzeko aukeria ta baliabidiak eskaintzearren.

I would also like to thank to **Sotiris Ioannidis** for giving me the opportinity to work with you and your team at the **Distributed Computing Systems Lab (FORTH-ICS)**. It was a wonderful experience and I could meet wonderful people and an amazing island. Gracias también a esos que hicisteis que disfrutara de las maravillas de **Creta**.

Zela ez, eskerrak departemantu kidieei be, emondako laguntziagaitxik eta atsedenalditxan deskonektatzia posible eitxearren. Bai abentura honen amaiera bitxartian egon zatenoi: **Aitor**[2], **Alain**, **Ane**, **Dani**, **Mikel M**, **Oscar**, **Pablo**, **Raúl**, **Unai**... Eta baitxa hasieran ikerketa munduan zirrikituak erakusten egon zinatenoi: **Aritz**, **Idoia**, **Iñaki**, **Joxe**, **Lorea**, **Lorena**, **Maite**... Eta bereziki **Mikel Iturbei** tesi osua irakurtziarren.

Gurasuei eta **Geaxiri** eskerrik beruenak, ni naizena banaiz, zuei esker naizelako. Baitxa gainontzeko **familixako guztixoi**, iraganian bidelagun izan baina egun urruti zagoztenoi eta noski, **lagunoi**, bizitzia disfrutatzeko dalako.

Azkenik, eskertza berezixa gailur honen azken txanpan sokalagun izan dodan **Janirei**, nahi bada, posible dala ikusaraztiarren.

Bat... bat ez da batera ailegatzen. Bi, bi t'erdi.

Contents

List of Figures

List of Tables

Acronyms

OSN *Online Social Network*

DDoS *Distributed Denial of Service*

ISP *Internet Service Provider*

NLP *Natural Language Processing*

SVM *Support Vector Machines*

CBR *Case-based Reasoning*

DNSBL *DNS-based Blackhole List*

SA *Sentiment Analysis*

KDD *Knowledge Discovery from Databases*

NOMENCLATURES

Those are the used nomenclatures in this dissertation:

Nomenclature	**Meaning**
DMNB	DMNBtext
BLR	Bayesian Logistic Regression
CNB	Complement Naive Bayes
NBM	Naive Bayes Multinomial
NBMU	Naive Bayes Multinomial Updatable
RT	Random Tree
RF	Random Forest
SVM	Support Vector Machine
ABM	Ada Boost with Naive Bayes
.c	*idft False, tft False, outwc True
.i.c	*idft True, tft False, outwc True
.i.t.c	*idft True, tft True, outwc True
.stwv	String to Word Vector
.go	General options
.wtok	Word Tokenizer
.ngtok	n-gram Tokenizer 1-3
.stemmer	Stemmer
.igain	Attribute selection using InfoGainAttributeEval

Table 1: Nomenclatures

**idft means Inverse Document Frequency (IDF) Transformation; tft means Term Frequency score (TF) Transformation; outwc counts the words occurrences.*

CHAPTER 1

INTRODUCTION

"Spam is an irrelevant or unsolicited message sent typically to a large number of users, for the purposes of advertising, phishing, spreading malware, etc." - Oxford Dictionaries

This chapter gives a description of the problems, and the main motivations that stimulate the author to provide a solution to the problems observed. The main objectives that had to be performed during the development of the thesis work are also given in this chapter, as well as the hypotheses that respond to the proposed objectives. Finally, the main contributions of this thesis are described, showing the publications associated with the main contributions.

1.1 Motivation

The mass mailing of unsolicited e-mails has been a real threat for years. Spam campaigns have been used both for the sale of products as well as online fraud. Researchers are investigating many approaches that try to minimize this type of malicious activity that reports billions of dollars of benefits in an underground economy.

Within the spam problem, most research and products focus on improving spam classification and filtering. According to Kaspersky Lab data, the average percentage of spam in email traffic in Q1 2016 amounted to 56.92%[1]. This per-

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

centage is 2.72 percentage point higher than in the previous quarter, Q3 2015[2], which demonstrates that spam is a current threat.

The same study shows a dramatical increase of spam containing malicious attachments in Q1 of 2016. This makes spam even more dangerous due to a gradual criminalization of it, confirmed by this growth. Several issues like social engineering, different types of attachments, diversity of languages take spam to a new level of danger.

Moreover, the current massive publication of private information in OSNs, give the attackers the possibility of using every single information against the users, like personalizing spam emails. Those sites are also becoming an attractive segment to act inside them. This is a significant risk if we take into account the amount of users that the most popular OSNs count: Facebook reached 1.65 billion monthly active users as of March 31, 2016 [3]; Youtube has counted over a billion users in 2016 [4]; and Twitter has 310 million monthly active users as of March 31, 2016[5].

Furthermore, in the same way that smartphones and online social networks are growing up, short messages traffic is increasing in all sorts of communication. 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 [4]. In the same way, WhatsApp, which is one of the most famous instant messaging applications, reached 1 billion users on February 2016[6].

Malicious campaigns in SMS communication systems are specially effective due to the phenomenal opening rate of 98% (for instance, email marketing reports a 22% opening rate)[7]. This demonstrates that there are billions of users whose privacy can be threatened 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 being reported as an emerging problem, specially in the Middle East and Asia [2].

1.2 Research Statement

This thesis aims to deal with the new business models that are growing around spam. The study has been done under the assumption that personalized spam messages will arise, specially inside OSN systems. Spammers might use personal information from OSNs for different purposes, such as to perform targeted attacks.

2016/

[2]https://cdn.securelist.com/files/2015/11/Q3-2015_Spam-report_final_EN.pdf

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

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

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

[6]https://blog.whatsapp.com/616/One-billion/

[7]http://goo.gl/CaxweY

The same way advertisement market targets users based on their behaviour and digital fingerprints they generate, malicious users can adopt similar strategies.

Moreover, despite the necessity of using millions of messages to capture a few potential buyers, the attackers can minimize the efforts to obtain higher profits. This is possible because they can focus the campaign only on users susceptible to fall into the trap.

The main objective of this thesis is to demonstrate that it is possible to develop new spam types using OSN public information, and also to design novel filtering model that would help with the detection of the former ones. This model should be able to improve spam filtering rates, thus detecting the intention behind the messages; intention of selling a product, intention of alarming users, intention of fooling users, intention of altering the concept of a product or service, etc.

1.2.1 Objectives

Bellow the most important objectives of this project are presented:

- To demonstrate that developing personalized spam is possible and a real threat.
- To improve current spam filtering techniques, by means of content-based analysis of text messages, aiding to detect the intentionality of the spammers.
- To demonstrate that above-mentioned techniques can be used to distinguish spam both on email and also on OSN and short message services.

1.2.2 Hypotheses

The next hypotheses try to formulate the research questions, that after giving a response and validation, will make accomplish the aforementioned objectives:

- Emails or texts that use information from the person that is targeted, attract the attention of the user, and therefore the personalized spam can maximize the click-through rate.
- Current Natural Language Processing techniques can help analyzing the content of the messages, thus giving new means to identify the intention of the spam text.
- The identification of the messages' intention can be approached by sentiment analysis and personality recognition techniques. These techniques will provide new features that improve current spam classification techniques.

- OSN content and short messages reduce the length of the text, but maintain the meaning of it. Hence, NLP and previous techniques will improve spam detection in these communication channels.

1.3 Contributions

We summarize the main contributions of this thesis, showing the publications associated with the main contributions:

- We evaluate the effectiveness of personalized spam campaigns. We demonstrate that a classic spam model using online social network information can obtain a 7.62% of click-through rate. We collect email addresses from the Internet, harvest email owner information using their public social network profile data, and analyze the response of personalized spam sent to users according to their profile. Finally we demonstrate the effectiveness of these profile-based emails comparing results between typical and personalized spam. [47] [45]

- We improve classification rates of current spam filtering techniques using sentiment analysis. We provide means to validate the assumption that being spam a commercial communication, the semantics of its contents are usually shaped with a positive meaning. We produce the polarity score of each message using sentiment classifiers, and then we compare spam filtering classifiers with and without the polarity score in terms of accuracy and the number of false positives. We demonstrate that sentiment analysis helps in three different types of spam filtering: email spam, SMS spam and OSNs spam. [43] [49]

- We demonstrate that it is possible to improve spam filtering using personality recognition techniques. Using publicly available labeled (spam/legitimate) datasets, we apply personality recognition techniques to each text and aggregate the personality feature to the original dataset, creating a new one. We compare the results of the best classifiers and filters over the different datasets (with and without personality) in order to demonstrate the influence of the personality. Experiments show that personality feature helps in email spam, SMS spam and OSNs spam filtering, improving the accuracy and reducing the number of false positive of the best classifiers. [46] [44] [48]

- We create a new spam detection method that combining sentiment analysis and personality recognition techniques is capable to detect non evident intent in spam texts. Once again, the validation of the method has been done using three different types of spam: email, SMS and OSN spam. [46]

1.4 Publications

Parts of the work of this thesis has been published or accepted for publication in national and international refereed conferences and journals:

Journal papers:

- Ezpeleta E., Zurutuza U., Gómez Hidalgo J.M., 2016. *A study of the personalization of spam content using Facebook public information.* Logic Journal of IGPL. *In Press.*

- Ezpeleta E., Zurutuza U., Gómez Hidalgo J.M., 2016. *Using Personality Recognition Techniques to Improve Bayesian Spam Filtering.* Journal Procesamiento del Lenguaje Natural, n 57. *In Press.*

Conference papers:

- Ezpeleta E., Zurutuza U., Gómez Hidalgo J.M., 2015. *An Analysis of the Effectiveness of Personalized Spam Using Online Social Network Public Information.* International Joint Conference - CISIS'15 and ICEUTE'15. Burgos, Spain, 15-17 June. Advances in Intelligent Systems and Computing 369, pp. 497-506, Springer International Publishing.

- Ezpeleta E., Zurutuza U., Gómez Hidalgo J.M., 2016. *Does Sentiment Analysis Help in Bayesian Spam Filtering?.* Hybrid Artificial Intelligent Systems. Seville, Spain. 18-20 April. Lecture Notes in Computer Science, Volume 9648, pp. 79-90, Springer International Publishing.

- Ezpeleta E., Zurutuza U., Gómez Hidalgo J.M., 2016. *Short Messages Spam Filtering Using Personality Recognition.* CERI '16. 4th Spanish Conference in Information Retrieval. Granada, Spain. 14-16 June. ACM International Conference Proceeding Series (ICPS).

- Ezpeleta E., Zurutuza U., Gómez Hidalgo J.M., 2016. *Short Messages Spam Filtering Using Sentiment Analysis.* Text, Speech and Dialogue TSD. Brno, Czech Republic. 12-16 September.Lecture Notes in Computer Science. *Accepted for publication.*

- Ezpeleta E., Zurutuza U., Gómez Hidalgo J.M., 2016. *Los spammers no piensan: usando reconocimiento de personalidad para el filtrado de spam en mensajes cortos.* RECSI 2016. XIV Reunión Española sobre Criptología y Seguridad de la Información. Menorca, Spain. 26-28 October. *Accepted for publication.*

1.5 Organization of Dissertation

The rest of this dissertation is organized as follows. First of all, some background information about spam is provided in Chapter 2. Different types of spam, as well as the current spam detection methods are explained. Finally, the problem caused by this threat in OSNs is presented.

In Chapter 3, personalized spam is created using OSNs public information in order to demonstrate the effectiveness of this type of spam. A system to automatically send spam campaigns is created and real experiments are carried out to analyze the differences between common spam versus personalized campaigns.

Chapter 4 presents a new spam detection method using the polarity of each text, which is extracted using sentiment analysis techniques. This method is validated using several datasets that represent different types of spam.

In Chapter 5, the new proposed spam filtering method uses personality recognition techniques in order to extract the personality feature of each text. Also, in this case, different tests are presented to validate the proposed method.

In Chapter 6, we explore how the combination of the two previous methods affect spam filtering. Each text is analyzed using sentiment analysis and personality recognition techniques, and the results are compared with the previously obtained results.

Finally, Chapter 7 summarizes the results and the contributions of this dissertation, and stimulating future work is discussed.

CHAPTER 2

TECHNICAL BACKGROUND

The goal of this chapter is to present the state of the art in the areas where the research project aims at contributing with new proposals. Additionally, it seeks to familiarize the reader with the terms used throughout the document.

This chapter has four different sections. First, the section where the spam threat is described. Second, learning-based filters for spam detection are explained. Third, content analysis based Natural Language Processing techniques for spam filtering are described. And finally, security issues in Online Social Networks (OSN) are presented.

2.1 Spam

Spam is one of the most common problem of digital communications. Currently, it can be found in several format such as email spam, SMS spam, social spam, opinion spam, etc. In this section a brief introduction to the spam problem is presented.

2.1.1 Email spam

56.92% of all emails sent worldwide are unsolicited emails (spam), online fraud (scam) and phishing emails [87].

Spam is defined as unsolicited emails sent via Internet. The term spam is taken from a luncheon meat, used in a Monty Python sketch in which Spam is included in

almost every dish. Internet users stated to use this word to denominate unsolicited emails, and currently this word is commonly accepted.

In SpamHaus's opinion [127], the word spam, is referred to unsolicited, massively-sent electronic messages. It includes: messages with commercial information, with attached malware, phishing scams, email chains or hoaxes. This features are required for marking an email as spam. From a technical point of view, emails with the following features are considered spam [127]:

- The recipient's personal identity and context are irrelevant because the message is potentially applicable to a large number of recipients.
- The sender does not have permission from the receiver to send him a message.

The transmission of spam has different negative consequences [27]:

- *Direct consequences:* Spam provides to the attackers a channel to sell products, to install viruses, to use the computer of the victim with other fraudulent purposes...
- *Network resource consumption:* As spam represents more than 50% of all email traffic, the bandwidth and storage consumption is significant and can affect communication infrastructures.
- *Human resource consumption:* It is a waste of time to sort through an inbox full of spam emails.
- *Lost email:* Spam detection techniques can mistakenly classify a legitimate message as spam.

Spam is considered an asymmetric threat. On the one hand, it is very easy to generate and send messages massively, but on the other hand, it requires a good organization and high costs for removing the unsolicited messages.

In order to mitigate the increase of these malicious practices, in recent years several legislative measures have been gradually adopted in the USA and Europe. However, the effectiveness of the laws have been proven very limited as users continue receiving spam emails [87]. Currently, different type of filters and techniques are used to classify spam messages automatically.

The differences between spam and legitimate messages are not directly reflected by simple attributes of the message such as sender's email address, the domain that is used for sending the message, its size, the existence of a given term, etc. However, there are statistical relationships between these attributes and the nature of the messages. Current techniques for email filtering are based on those statistics.

2.1.2 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. Among all the instant message services, in this dissertation we are focusing specially on SMS messages. Those are structurally similar to other currently very consumed short message applications such as WhatsApp or even Twitter. Our decision to focus on certain types of messages is principally based on the public access to labeled datasets needed to generate and validate classification models. This approach 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 large population[1], and also used by people of countries where SMS services are not charged by mobile operators.

In [32], authors presented a survey on filtering SMS spam and showed recent developments in SMS spam filtering. Also, they show a brief discussion about publicly available corpus and availability for future research in the area.

Authors in [7] compared different machine learning methods and indicated that Support Vector Machine technique was the best one during their experiments. They obtained an accuracy of 97.64% in spam filtering using this method. Furthermore, they offer a public and non-encoded SMS spam collection that can be used by the research community. This study brings us the possibility to test with the same dataset and to compare results.

In other recent studies such as [111] and [109], two-level classifiers, where two classifiers are applied and the second level classifier complements the first level, are used to improve the classification. In this study we are going to focus on improving one-level learning-based classifiers.

2.1.3 Spam detection

The first known spam message appeared publicly on 3^{rd} May, 1978, when DEC Marketing Company sent a message to 400 people via Arpanet [151, 15]. Since then, spam has grown in an underground industry sending out billions of messages every day [86]. The user needed to filter and delete those unsolicited messages manually. Therefore researches started to design and develop automatic filters as a premier solution, combining these filters with reputation, and protocol related techniques. Currently it is possible to find several types of filters, which can be classified into two groups: content-based techniques and collaborative techniques [117] [106].

[1] https://goo.gl/g6R7uW

Content-based techniques

Content-based techniques are based on the analysis of text or images of messages to classify them. Some commonly used features are: the number of words, the average length of words, keywords, etc. As authors explain in [140], these filters can be hand-made rules, also known as heuristic filters, or learned rules using Machine Learning algorithms. There are several different types of content-based filters as authors described in [106]:

- *Content-based filtering using Bayes filters:* Probabilistic classifiers based on Bayes' theorem, they assume strong independence between features. Naives Bayes and Flexible Bayes are two of those filters. More details about those filters are presented in Section 2.2.

- *Classification using Support Vector Machines(SVM):* A SVM constructs a hyperplane that separates spam and legitimate classes [106].

- *Combinations of weak classifiers:* Several weak classifiers are used (generally decision trees) to improve the classification in a cooperative way. AdaBoost and Random Forests are two examples of this type of classifiers [106].

- *Rule learning algorithms:* The knowledge is stored as rules. Those rules are extracted from an email set. RIPPER and IREP [25] models can be good examples of rule learning algorithms.

- *Pattern recognition techniques:* The main objective is to obtain patterns analysing exhaustively the body of the messages. Chung Kwei technique [133] is a well known example.

- *Term and document frequency analysis:* They are based on the analysis of the frequency of documents containing a term, and the frequency of a term in a corpus. One example of this type os content-based filter is Rocchio [83].

- *Case-based reasoning(CBR):* CBR's are Hybrid models. Past experiences are used to solve new problems. CBR implements four phases for each classification: recovery of similar cases, reuse, revision and learning. Popular examples of those model is ECUE [33].

- *URD - URL Semantic analysis:* First, URLs are extracted from the body of the received message. In the next step they execute a content-based analysis in the text of the URL. This technique makes possible the detection of spam in URL's that are not categorized, but have very descriptive text [106].

- *URAC - Content analysis of URLs:* In this case, the URL contained in the body of the message is extracted. The HTML code of the web site is obtained. Then, such code is analyzed in order to detect spam using different dictionaries. This technique offers: cache support, redirections support, analysis of the redirection URLs and analysis of the URL extracted from the HTML code [106].

Collaborative techniques

Collaborative techniques are based on the work of the whole community, as information about spam messages is shared among the users, instead of using individual filters. If one user receives a spam message and manually filters it, all users in the community get information about such spam. This technique minimizes false positive rates.

Those techniques are classified in [106] in the next groups, based on the information that is shared:

- *Content-summary based filters:* The hash (MD5, SHA1, etc.) of the message content is shared on those filters. Some examples are: Razor [125] and Pyzor [5].

- *Header extracted data-based filters:* Data of the sender (email address, domain or IP address) or the subject is shared. There are two main types: black list based and white list based. Well known black list example is SBL [126].

- *DNS-based Blackhole List (DNSBL) - Spammer catalogue:* IP addresses are analyzed using DNSBL providers. First, a list of IPs related with an email is built. After that, searches are executed using the DNSBL providers.

Reputation-based techniques

Reputation-based techniques store different information of the spam sources, such as IP address or the history of emails. Some of those techniques are [1]:

- *LRL - Contextual reputation database:* LRL technique aims at reducing the percentage of false positives in the filter. The history of the sender is used (analysing sent emails) to classify new messages.

- *LNI - Optenet reputation database:* In this technique the IP addresses of the emails are analysed. The sender IP address is searched in the reputation

database that Optenet company [2] owns. If the email is classified as spam, a reputation analysis of the email elements is executed.

- *URL - Verification of URLs in category database:* This technique work in the following way: First, all the URLs are extracted from the incoming message. After that, those URLs are stored in one list. In the next step, the system attempts to find those URL in a database to obtain its class. If there is a coincidence between this class and any of the classes configured for URL detection, the message is classified as spam.

- *IPR - External reputation database:* Anti-spam services are able to combine with external IP reputation services, such as Commtouch[3]. This technique analyzes IP addresses extracted from the message, and calculates the reputation using the information of an external supplier.

Other technique:

Some other techniques for spam detection have been proposed that can not be classified into one of the analysis types we have considered so far. These techniques are described below:

- *SPF - Sender Policy Framework:* This technique analyzes the IP address of the message. Specifically, it verifies if the sender's IP address is allowed to send a message using the domain name that uses.

- *LNG - Shared lists of suspicious addresses:* A database with usual spam sender addresses is used to classify messages. If the sender address is in the database, the message is classified as spam.

- *MD5 - Proprietary digital signature verification:* This technique uses the MD5 hash of the attached file in order to search for such files in a historical database of spam attachments.

- *RDNS - Reverse DNS Resolution:* This technique is used in the cases were the IP address is inserted in links that are in the content of the message. For this purpose, the URL associated to the IP address is calculated, and then it is searched in a database.

[2] http://www.optenet.com/
[3] http://www.commtouch.com/

2.2 Learning-Based Filtering

In the literature we can see that in many cases Bayesian Filters are used instead of learning-based filters.

During the last years, several techniques to detect unsolicited emails have been developed [136]. Among all proposed automatic classifying techniques, machine learning algorithms have achieved more success [27]. For instance, different studies such as [153] obtained precisions up to 94.4% using this kind of techniques.

In this thesis we focus on filters that are able to work with the content of the messages: content-based filters.

Teli et al. presented in [141] a comparison between various existing spam detection methods including rule-based system, IP blacklist, Heuristic-based filters, Bayesian network-based filters, white list and DNS black holes. They concluded that the most effective, accurate, and reliable spam detection methods were the Bayesian based filters.

In [103] some of the content-based filtering techniques are studied and analyzed, and the Bayesian method was selected as the most effective one (classifying correctly the 96.5 % of messages). Furthermore, in [38] authors demonstrated that although more sophisticated methods have been implemented, Bayesian methods of text classification are still useful.

2.2.1 Learning Algorithms

The learning algorithms behind learning-based spam filtering techniques are considered the central part according to [140]. These algorithms aim at obtaining the most accurate approximation to a perfect classification (no mistakenly classified messages). In spam classification several learning algorithm families have been applied: probabilistic Naive Bayes [8, 9, 62, 121, 128, 138], rule learners [50, 121], Decision Trees [21, 50], linear SVM [37], classifier committees [138], Cost-Sensitive learning [71], and Instance Based k-Nearest Neighbors (kNN) [9]. In this section we explain the most important learning families.

Probabilistic Approaches

Thanks to its simplicity and proven results, those approaches are commonly used in spam filtering. They are one of the first presented filters, but are very used in recent years [61]. Based on the Bayes Theorem [94], they compute the probability for a document d to belong to a category c_k as authors explain in [140]:

$$P(c_k|d) = \frac{P(c_k)P(d|c_k)}{P(d)}$$

In some cases, it is common that using Bayesian learner terms in a document are considered independent and their order is considered irrelevant. That is the way "Naive Bayes" learner is defined [94, 105], being the version published in [61] one of the most frecuently used.

Decision Trees

Decision tree learning uses a decision tree as a predictive model which define different observations about an item in order to classify it.

A Decision Tree is a finite tree with branches annotated with tests, and leaves being categories. It uses a Boolean expression about the items in the document. To classify a text, we start from the top of the tree and taking into account the different conditions in branches we follow true answers through the tree. We repeat the evaluations until a leaf is reached and a text is classified.

Some of the most common algorithms are: ID3 [54], C4.5 [26], and C5 [95].

Rule Learners

The base of Rule Learners learners are the "if-then" type rules, which are designed to be applied sequentially.

Like decision trees, rule learning algorithms are popular because the knowledge representation is very easy to interpret.

Support Vector Machines

Support Vector Machines are currently considered as one of the best algorithms in spam classification [28, 37].

A SVM is a discriminative classifier formally defined by a separating hyperplane. The algorithm outputs an optimal hyperplane, which categorizes new examples, given a labeled training data.

k-Nearest Neighbors

New instances are classified as a majority class among the nearest k neighbors among all the training data.

Those kind of algorithms are called lazy, because during the training phase only the instances are stored, and they do not build any model. The classification is done once the test instance is evaluated. It is a non parametrical algorithm, because no assumptions are done about the distribution of the data.

The popular kNN classifier was introduced in [170].

Classifier Ensemble

Classifier Ensemble learning methods generate multiple classifiers to form a committee by repeated application of a single base learning algorithm. Three main techniques are used: bagging, boosting, and stacking.

In bagging, several classifiers are trained using different subsets and at the end, these classifiers vote a final decision.

In boosting, a series of classifiers are trained on the dataset. Those classifiers put more emphasis on previously failed training examples. In other words, each classifier makes more effort in the failed training set of the previous classifier in order to improve the classification in a series of classifiers.

Stacking is a short name for Stacked Generalization [138] which uses multiple models to combine their predictions.

Cost-Sensitive Learning

Cost-Sensitive Learning takes into consideration the cost of misclassification, in order to minimize the total cost. For example, in spam classification, it is more important to reduce the number of false positives than true positives. This type of learning give the posibility to penalize more one case than the other.

Three main methods for making algorithms cost-sensitive are compared in [71], concluding that the most effective one is the combination of SVMs and weighting.

2.3 Natural Language Processing Techniques for Spam Filtering

In [96], Natural Language Processing or NLP is defined as:

Definition 2.3.1 *A theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.*

NLP is used for several applications. In [23], the most important disciplines are summarized as:

- *Natural Language Text Processing Systems:* The objective is to translate potentially ambiguous texts into unambiguous representations to perform matching and retrieval using them [96].

- *Abstracting:* Abstracting techniques aim at generating summaries of texts.

- *Information Extraction:* It is possible to extract certain information using NLP.
- *Information Retrieval:* This is concerned with storing, searching and retrieving information.
- *Machine Translation:* Automatically translate text from one language to another.
- *Sentiment analysis:* Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.

Currently, the most known application of NLP is machine translation, and it is mainly used for website translation. But those techniques make a literal translation of the sentences, while the meaning may vary depending on the context of the message. In human language, the relation between the meaning and the context is very important.

As mentioned in [55], in the late 90's some studies started analysing those relations (topic analysis) [11, 88]. But it was in the beginning of the year 2000 when the research effort grew up exponentially [72, 135, 99, 34, 147, 132].

Since 2000, NLP approaches are used for sense detection in combination with Machine Learning and semantic analysis techniques to improve the results.

Focusing on how NLP works, 4 stages can be defined [55].

- *Morphological analysis:* Morphology is the structure of the word. It aims to analyze the individual words into their components, and it is able to detect the relation between the minimum units that form the sentence [10]. For this purpose, in this type of analysis it is important to separate non-word tokens from words. Moreover, the analysis is concerned with inflection and also with derivation of new words from existing ones. One example of this kind of analysis is Stemming, which extracts the root of each word.
- *Syntactic analysis:* Here, the analysis focuses on the words that form a sentence to know the grammatical structure of the sentence [139]. The words are transformed into structures that show how the words relate to each others. These grammatical units are formed by rule-sets like: (1) PP = prepositional phrase; (2) NP = noun phrase; (3) VP = verbal phrase and (4) Det = determinant.
- *Semantic analysis:* [60] This analysis examines sentences and its codification to get the meaning and sense according to the context. This stage uses the meanings of the words to extend and disambiguate the result returned by the syntactic parse.

- *Pragmatic analysis:* This is an additional stage of analysis concerned with the pragmatic use of the language [167]. This is important in the understanding of texts and dialogues. This analysis is one of the most complex because it tries to contribute with more significant information about word senses, according to speech and participant information [55].

Spam detection using NLP

According to [59] it is possible to improve spam detection using NLP techniques. Giyanani and Desai used NLP for the design of a new method. Their method blocks spam messages based on the sender and the content of the email.

2.3.1 Text Mining

As Tan described in [148], text mining is also known as text data mining [67] or knowledge discovery from textual databases [52]. It is a process of extracting interesting and non-trivial patterns or knowledge from text documents. In studies such as [51, 143], it is considered as a subarea of data mining or knowledge discovery from databases (KDD).

Hearst [66] defines Text Mining as:

Definition 2.3.2 *The discovery of new, previously unknown information, by automatically extracting information from different written resources.*

According to [148], text mining is used in different research areas as Information Retrieval [68], NLP [90] or Information Extraction [165]. It is considered an interdisciplinary field which is composed by: data mining, web mining, information retrieval, information extraction and NLP [130].

In order to apply text mining techniques text preprocessing is applied usually previously, specially for mining large document collections. In this process, a set of words is collected from the text documents and stored in a dictionary. Filtering, lemmatization and stemming methods can be used to reduce the size and complexity of the document [73].

- *Filtering:* To remove words that are not related to the documents (articles, conjunctions...).

- *Lemmatization:* To group declined forms of words, to be analyzed as a single term.

- *Stemming:* To identify each word by its root (For example: removing "ed" from past tense).

The main reason for applying text mining techniques is to structure text documents [73]. The following are the most important techniques in text mining [130].

- *Feature extraction:* The main task is to extract parts of the text and assign specific attributes to them.
- *Classification:* Text classification aims to assign pre-defined classes to text documents [107].
- *Clustering:* The goal is to find groups of documents based on their own features, without pre-defined classes.
- *Summarization:* It is the operation that decreases the amount of text in a document, minimizing the loose of its meaning.

According to [130], text mining has been applied to text areas:

- *Internet:* To support automatic classification of texts.
- *Insurance:* Insurance companies receive information that is stored and can be given as input to text mining algorithms to obtain useful knowledge.
- *Sentiment analysis:* To identify how sentiments are expressed in texts and whether the expressions indicate positive, negative or neutral opinions about the topic.
- *Bioinformatics:* To translate biomedical literature from unstructured format into a structured format.
- *Identifying patterns and trends in journals and proceedings:* To obtain knowledge from a big amount of journals and articles to be used by researchers.

Besides these applications, in [39] authors demonstrate that it is possible to develop an anti-spam system using text mining techniques. Further, in [93], a novel text mining model is developed and integrated into a semantic language model for the detection of spam reviews.

2.3.2 Sentiment analysis and opinion mining

In [98], sentiment analysis (SA) or opinion mining is defined as

Definition 2.3.3 *The computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes.*

Text analysis and computational linguistics are used in SA to identify and extract subjective information in source material natural language processing.

As explained in [119], the area of SA has had a huge burst of research activity during the last years, but there has been a continued interest for a while. It is possible to find early works in this topic, like those in [19] and [166]. Later, research was focused on topics like interpretation of metaphors, narrative, point of view, affect, evidentiality in text, and related areas [65, 75, 85, 137, 159, 161, 162, 163, 164]. But the huge increase of SA-related research effort begins in year 2001, when researches started to solve different problems using SA techniques [20, 30, 31, 36, 99, 120, 150, 152, 158, 171].

Currently there are several research topics on opinion mining [98]:

- *Document sentiment classification:* The main objective of this area is classifying the opinion of a document as positive or negative [119]. In order to classify such sentiment, some researchers use supervised learning techniques, where three classes are previously defined (positive, negative and neutral) [120]. Some other authors propose the use of unsupervised learning. In unsupervised techniques, opinion words or phrases are the dominating indicators for sentiment classification [154].

- *Sentence subjectivity* and *sentiment classification:* In the same way that it is possible to classify documents based on their polarity, it is also possible to classify sentences. Given a sentence, two sub-tasks are performed: (1) Subjectivity classification [135, 134, 160], to define if it is a subjective or objective sentence, and (2) sentence-level sentiment classification (positive or negative opinions).

- *Opinion lexicon expansion:* In the previous tasks, the use of opinion related words is necessary. In this area, those words are generated from three different approaches: (1) manual approach, (2) dictionary-based approach, using small, manually made, set of word combinations with online dictionaries [74, 89], and (3) corpus-based approach, rely on syntactic or co-occurrence patterns and also a seed list of opinion words to find other opinion words in a large corpus [64].

- *Aspect-based SA:* There are some cases where the author writes both positive and negative aspects of an entity, although the general sentiment on the entity may be positive or negative. In those cases, document or sentence classification is not enough to extract the details needed for many applications. To obtain these detailed level, we need to focus on the aspect level, where two tasks are determinant:

  - *Aspect extraction:* This task aims to extract the aspect that have been evaluated. For example, in the sentence: "The consumption of this car is very low", the aspect is "consumption".
  - *Aspect sentiment classification:* It specifies if the opinion related to the aspect is positive, negative or neutral. In the previous example, the opinion about the "consumption" is positive.

- *Mining comparative opinions:* Analyzing comparative sentences as useful is as analyzing regular opinion sentences. In this case, the objective is to evaluate comparative sentences, where two different 'products' are compared. Different studies solve these type of problems [31, 35, 56, 79].

Opinion spam detection

In the Web 2.0 era, writing and reading user opinions about a product is very common. For instance, many people rely on opinions to decide if it is worth buying a product or not. Based on this argument, some companies started writing false opinions to increase their sales. This problem called opinion spam was introduced by Jindal and Liu in [81, 80].

It is possible to deal with opinion spam using different techniques as:

- *Spam detection based on supervised learning:* It uses machine learning algorithms to train a classification model based on previously labelled classes. While manual labelling is time consuming, previous research exploit other possibilities such as detection of duplicate reviews [81].

- *Spam detection based on abnormal behaviours:* Due the difficulty of correctly/reliably detect opinion spam based on supervised learning (because of false positive/false negative problems), researchers have proposed behaviour-based detection models. In [97] for example, authors identify unusual reviewer behaviour models so as to detect the spammers. They derive an aggregated behaviour scoring method to rank reviewers according to the degree they demonstrate spamming behaviours. In [82], authors identify unusual reviewer behaviour patterns via unexpected rule discovery [98].

- *Group spam detection:* This technique is focused on detecting a group of spammers that work together to promote or demote a product or a brand. The algorithm is explained in [108].

Opinion mining and OSNs

Due the increase of OSN popularity, they are becoming a very good scenario and information source for SA.

In [112], authors apply SA techniques over more than 1,000 Facebook posts about newscasts, comparing the sentiment for the Italian public broadcasting service RAI.

Moreover, in [149] Tan et al. demonstrated that the information about social relationships extracted from OSNs can be used to improve user-level SA.

2.4 Security in Online Social Networks

An Online Social Network (OSN) can be defined as:

Definition 2.4.1 *A web site whose purpose is to allow users to interact, communicate, share content and create communities.*

In [17], the authors explained the possible activities that a OSN is supposed to allow users:

- To construct a public or semi-public profiles within a bounded system.
- To articulate a list of other users with whom they share a connection.
- To view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site (Followers in Twitter, friends in Facebook...).

Over the last few years, OSNs have become one of the main ways to keep track and communicate with people. Sites such as Facebook and Twitter are consistently among the top 10 most-viewed web sites on Internet [6]. Moreover, statistics show that, on average, users spend more time on popular social networking than in any other sites [6].

The tremendous increase in popularity of OSNs allows malicious users or groups to collect a huge amount of personal information about users [41]. Unfortunately, this wealth of information, as well as the ease with which one can reach many users, also attracted the interest of attackers. Consequently, researches started to protect user privacy in OSN.

Online social network security

Considering the large number of users and information in OSNs, the protection of users' privacy, of their communication channels, as well as trust among parties becomes necessary. In consequence, OSNs spend a lot of resources generating and improving their security strategies. In [145], authors present a tool called Facebook Immune System. They believe this tool has contributed to making Facebook the

safest place on the Internet for people and their information. But OSNs provides attackers different possibilities to perform attacks. In [3], authors explained the logical evolution from traditional attacks to more specific forms that leverage OSN information.

For example as Mahmood shows in [100], OSNs can be used by criminals with different objectives:

- To find incentives for a crime.
- To plan the crime's execution.
- To make an escape plan.

In order to satisfy such objectives, criminals only need few information from the user profiles. As a result, protection of users is not straightforward, if they share their own information publicly.

Furthermore, Makridakis et al. [102] demonstrate that online social networking web sites have the ideal properties to become attack platforms. Their results showed that it is possible to implement an application that can:

- Launch DDoS attacks in OSNs.
- Retrieve remote files from a user's machine.
- Leak user's private data.

In [122] authors describe how can an OSN like Facebook be exploited and converted into an attack platform. They show how a Facebook application can be used to collect user information, and then perform malicious actions against them, based on user's installed applications and open ports. They also showed some security leaks they observed during the study.

Polakis et al. [124] focused on the security problem of the Social Authentication system used by Facebook. This system required users to identify some of their friends in randomly selected photos in order to get into the system. Authors used widely available face recognition software and services to break the authentication with high accuracy. Finally they developed an automated social authentication breaking system.

Focusing on Facebook, Bonneau et al. summarized in [16] the main vulnerabilities of OSNs, are (1) resources for social engineering, (2) access to personal data, (3) data centralization and (4) underground economy (OSNs earn money with our information). Thereby, the most feasible threats in Facebook are:

- Clickjacking. Hidden code in some malicious websites can give to the attackers the control of your browser. By clicking links to those pages, attackers can insert code or information in your Facebook profile.

- Specific worms developed to attack OSNs users.
- Spam in private messages, in the wall, in events and in chat.
- Malicious pages and groups. Where attackers attract users with interesting advertisements.
- Fake notification messages. Attackers send emails using the Facebook's style using public data.

To deal with those security problems in Facebook, many studies are being carried out. In [129], authors implemented a Facebook application with the objective of detecting all kind of malware distributed on the OSN. They designed an efficient malware detection method which takes advantage of the social context of posts. During the research, they could observe that the malware inside Facebook is evolving from the traditional spam to specific spam in applications and in pages hosted on Facebook.

Other example of study with the aim of protecting users from different types of attacks is [70]. The authors demonstrate the possibility of creating an online social honeynet to protect the users from web crawlers. They show that the amount of OSN data disclosed to the web crawler can be kept at low levels.

Online social network spam

Numerous research related with spam and OSNs has been carried out. Researchers from the University of California at Santa Barbara proved that spam is a very big issue for OSNs [146]. In their research, they created a large and diverse set of false profiles on three large social networking sites (Facebook, Twitter and MySpace), and they stored the contacts and messages they received. They then analyzed the collected data and identified anomalous behaviors of users who contacted their profiles. Based on the analysis of this behavior, they developed techniques to detect spammers inside OSNs, and they aggregated their messages in larger spam campaigns. Results show that it is possible to automatically identify accounts used by spammers, and block these spam profiles.

Other study can be found in [58]. Authors carried out a study to quantify and characterize spam campaigns launched from accounts on OSNs. They studied a large anonymized dataset of asynchronous "wall" messages between Facebook users. They analyzed all wall messages received by roughly 3.5 million Facebook users, and used a set of automated techniques to detect and characterize coordinated spam campaigns. This study was the first to quantify the extent of malicious content and compromised accounts in a large OSN. While they cannot determine how effective these posts are at soliciting user visits and spreading malware, their

result clearly showed that OSNs are now a major delivery platform targeted for spam and malware. In addition, their work demonstrates that automated detection techniques and heuristics can be successfully used to detect social spam.

Further, a spam detection framework is proposed by Wang et al. in [157]. They developed a framework for spam detection which is able to run across OSNs. This framework has the next characteristics that may benefit the users' security: (1) the spam detection system collects more data as it uses several OSNs; (2) the framework is scalable (it is possible to add new detecting techniques); and (3) new OSNs can be plugged into the system easily. An equally important study is presented in [40]. The authors developed a tool that detects compromised accounts based on anomalies detected in user behaviour. To build a behavioural profile seven features are used: Time (hour of day) the account is typically active, message source, message text (language), message topic, direct user interaction (interaction history for a user) and proximity. The tool detects sudden changes in user activities, being those changes malicious or benign. Then they look all the profiles that have shown similar changes within a short period of time. They detect the compromised accounts assuming that this similar behaviour changes are part of a malicious campaign.

In [24] and [57], authors used classification and clustering techniques to detect spam campaigns inside different OSNs such as Facebook and Twitter.

In terms of spam inside OSNs, it is important to mention that a huge amount of studies about spam in Twitter have been performed. As we can see in [92], this micro blogging website is very useful to use as information diffusion tools, hence it becomes a suitable platform to perform spam attacks. At the same time, the open nature of Twitter makes it a worthy platform to carry out various research works. For example, authors explain in [168] how criminal accounts mix into and survive in the whole Twitter space.

Moreover, Song et al. [144] explain how to detect spam accounts instead of detecting spam messages. They demonstrate how spammer detection without reading messages is possible. They use distance (when two users are directly connected by a single edge, the distance between the users is one) and connectivity (the strength of the relationships) between receiver and recipient which are hard to manipulate by spammers and effective to classify them.

In [156], authors explain how to detect spam accounts in Twitter, but in this case they use both graph-based and content-based detection methods. In the first one they use the relationship between number of users they follow and number of followers they have. Results show that few spam accounts follow large amount of users, even if some spammers have many followers. In this last case, they use novel content-based techniques such as duplicate Tweets, replies and mentions, HTTP links and trending topics for spammer detection. Later, Yang et al. [169] explain

how to improve previous spam detection system in Twitter. They were able to achieve a higher detection rate while keeping an even lower false positive rate.

While most of the research focus on spam campaigns that might appear inside OSNs [172], we still think that a combination of typical spam and OSN spam exposes serious threats that needs to be addressed.

CHAPTER 3

EFFECTIVENESS OF PERSONALIZED SPAM

With the rise of online social networks (OSNs), and more specifically Facebook, which has more than 1.59 billion monthly active users as of December 2015[1], the extraction of personal information that users leave public on their profiles multiply spam success possibilities. Facebook provides a great opportunity for attackers to personalize the spam, so a much lower volume of messages would get a higher return on investment.

The main objective of this chapter is to measure the consequences of displaying information publicly in OSNs. It also aims to demonstrate that techniques for generating personalized email that evade current spam detection systems while increasing the click-through rate can be developed. These techniques can enable new forms of attacks. First we extracted email addresses while crawling the Internet. These addresses were then checked on Facebook to look for related profiles. Once obtained a considerable quantity of user addresses, we extracted all the related public profile information and temporally stored it in a database. Then this information was analyzed in order to design user profiles based on their main activities in Facebook. Email templates were generated using common information patterns. Finally, to demonstrate the effectiveness of these templates when systems circumvent spam detection, different experiments have been performed. We collected sufficient evidence to confirm that the goal was successfully achieved.

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

3.1 Relevant Proposals

During the last years several works about the possibilities to create personalized spam or collect personal information from different OSNs have been proposed. For instance, in [63] authors launch targeted and non-targeted attacks on different channels using information from Facebook accounts.

In [16], researchers at University of Cambridge and Microsoft analyzed the difficulty of extracting user information from Facebook to create user profiles. They described different ways of collecting user related data, and they demonstrated the efficiency of the proposed methods. Authors conclude that the protection of Facebook against information crawlers was low. They also proved that large scale collection of data is possible. While it is true that Facebook has improved its systems' security since then, like limiting its own query language, the research proved that data extraction was effective.

In [14] researchers found a Facebook vulnerability giving attackers the possibility of searching people through email addresses in OSNs. Starting from a list of different emails, they managed to connect those email addresses with the account of their owners. After that, they collected all the information they could, and created different user profiles. This work gave a baseline for allowing attackers to launch sophisticated and specific attacks, but still did not realize about the potential of creating personalized spam campaigns. In the same direction, Polakis et al. showed in [123] the potential of creating personalized spam campaigns in different OSNs.

3.2 Creating a Personalized Spam Campaign

As shown in Figure 3.1, our study followed four different phases. First, we collected a large amount of public information from Facebook. To do this, we used email addresses that were publicly available when crawling the Internet. Then we computed a number of interesting statistics from the collected information that will be shown later. As a result of the data analysis, different user profiles were identified, and used them as customizable email templates. Once we had defined these templates, we developed an automatic email sending system and conducted two different experiments. Finally we analyzed the results obtained in the experiments.

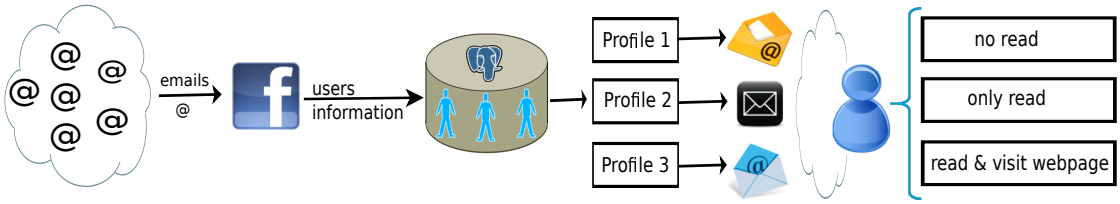

Figure 3.1: Full process of personalized spam campaign creation

3.2.1 Collection of data

This process has been performed in three steps:

Email address harvesting

In this task we considered two options: the first one, obtaining the email addresses using the techniques explained in [123], where they get e-mail addresses using various combinations of public information from OSN users. The second, using publicly available applications that automatically harvest email addresses from simple search queries over known search engines. The one used by the authors[2], generates a query for the search engine using a given keyword, and extracts email address patterns from the search result. We used a set of common keyword patterns such as "facebook", "hotmail", "gmail", "yahoo", "msn", and used both Google, Ask and Yahoo as search engines. Those patterns harvest email addresses from popular email service sites, which are at the end commonly used as user related data for online social networks.

Email address validation

Facebook offers the option to find the profile that is associated to a given email address. we developed an application that first authenticates a user to the OSN, and then searchers for a user corresponding to each email address harvested before.

Once the user profile was found, we extracted and saved the user's ID and their full name.

Next URL is used to check if a given email (changing the word "EMAIL" with the email addresses) corresponds to a specific Facebook user.

```
http://www.facebook.com/search.php?init=s%3Aemail&q=EMAIL&type=users
```

[2]http://www.fast-email-extractor.com/

Collection of the information.

Facebook allows extracting information from the source code of all its web site pages. In order to do it, it is mandatory to access directly to the page from which we want to extract the information. Therefore, in this program we have used a user identifier from Facebook to connect directly to the user information page. Thus, we have visited all user pages and we have been able to extract all the public information that users have in the Facebook database.

Below is the address where all public information of each user can be found. 'USERUID' correspond to the ID that Facebook gives each user, which is stored in the database.

`http://www.facebook.com/profile.php?id=USERUID&v=info`

Results

We found that 19% of the collected email addresses have a corresponding Facebook account associated to it. We found 22,654 Facebook accounts using 119,012 email addresses (19.04%).

3.2.2 Data processing

At this stage the aim has been to treat the data stored in the database to extract user profiles. We have summarized the most useful information about each user, which is related to the topics presented in Table 3.1. Most of the information stored is numeric, namely, the number of sports played by a person or number of activities that users entered in their own profile. Additionally each user also has logical or *boolean* data types, which indicates for example if the user in question is a male or not.

Summary					
id	bigint	tv	int	religion	boolean
sport	int	game	int	politic	boolean
team	int	activity	int	man	boolean
sportman	int	interest	int	woman	boolean
music	int	studies	int	partner	boolean
book	int	languages	int	company	boolean
movie	int				

Table 3.1: Design of the statistical table

Using those user profile-based features, we gathered interests and user-related attributes to generate a set of statistics that could describe the behaviour of OSN

profiles.

To obtain the more representative variables, we created a tag-cloud where we use every variable introduced by user, were appearing frequency increments the variable name.

Figure 3.2: Tag-cloud

Figure 3.2 shows that the most commons variables are Music, Gender and Studies. Table 3.2 gives a detailed view of those.

Variable	Amount	Percentage
man	8,786	39%
woman	6,189	27%
music	5,788	26%
titles	5,612	25%
company	5,149	23%

Table 3.2: Number of users who have entered each variable (total users: 22,654)

With the extracted statistics, we can draw the following conclusions:

- As a result of descriptive analysis of the collected data, we found that 25.21% of the users do not insert any type of personal information, and 82,25% of the users have entered 3 or less types of variables.
- 66% of the collected users leave their gender public.
- Taking into account only the users that have at least one public variable, those are the percentage of the most common variables: gender 88%, music 34%, studies 33% and company 30%.

- The variables related to personal information that are most added by the users (gender, music, studies and company), are still in very low ranges as to be processed and used for clustering user profiles, as shown in Table 3.2.

3.2.3 Personalized spam

The objective of this phase is to create different email templates that will be used later. With this templates, it is possible to send personalized mail to all Facebook users stored in the database. Once the templates were designed and implemented, a strategy to count the number of users that "bite the hook" of the spam was designed. For this we have implemented a website.

Mail templates

Before any other action, the first step was to define a template through which we were able to send personalized emails to the people.

As shown in Table 3.2, the most abundant variables are those related to the gender. Although these data cannot be used for creating templates, it can be used for implementing a formal greeting according to their gender. Following the process defined in Figure 3.3, we have used the other three most common variables to create spam templates. That is, if a user has entered his favourite music group in the profile, it will receive a personalized **music template**. However, if the person has not added any singer or group in Facebook and has added the university in which he or she has studied, she will receive a personalized message with the **studies template**. And if none of those two had been added but the information refers to their current job or a company in which the user works, she will receive a personalized the email using the **company template**.

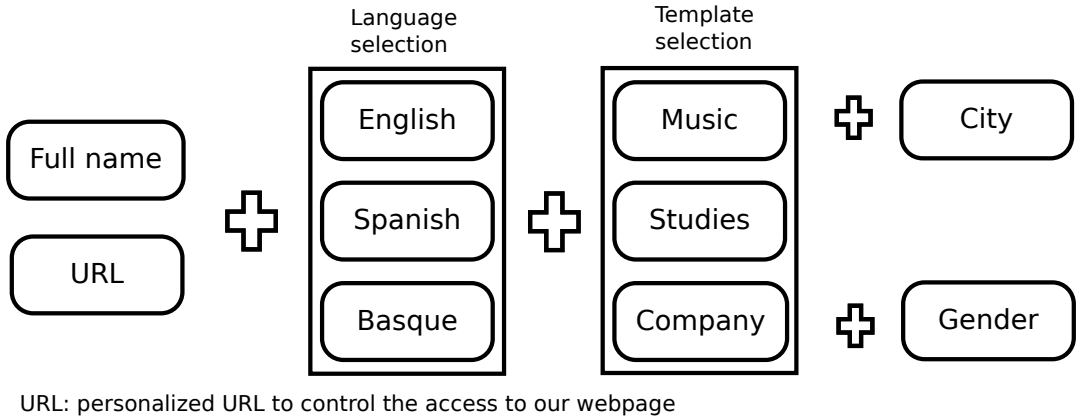

Figure 3.3: Email templates

For better customization, we have also used some profile fields such as the language, the name of each user, the gender, and the city in some of the templates. Each template includes a customizable URL that will track the action made by the spam recipient.

Templates: In the following lines, the fixed text of the different templates are presented:

- *Typical spam:*
  - Subject: "Amy Smith needs your help".
  - Content: "Amy Smith is a 10 year old girl who has a serious illness and she needs your help. You can read her story here *(URL)*.".
- *Music template:*
  - Subject: "New tour of *(the name of the favorite music group)*".
  - Content: "Hi *(name)*!

    Do you know that *(music group)* is preparing a new tour?

    And the most important think: They're thinking of giving a concert in *(city)*!

    Do you want more information? You can read more about the tour here *(URL)*.".
- *Studies template:*
  - Subject: "*(university or school name)* students and former students party".
  - Content: "Hi *(name)*!

    We are a student group from *(university)* and we want to inform you that we are organizing a party with students and former students.

    You can read more information about the party and sign up here *(URL)*.".
- *Company template:*
  - Subject: "*(name of the company)* employees and former employees brunch".
  - Content: "Dear Mrs/Mr *(name)*,

    With this email, *(company)* wants to invite you to the employees and former employees brunch.

    You can read more details and give your opinion here *(URL)*.

    Sincerely yours, Director of *(company)*.".

Website

Access to this site should only come from the users personal email. The website is defined to store information about which user, and from which specific spam message has reached the site. Considering all these details, we decided that the most appropriate way was to introduce parameters in the URL which will be included in emails. When the user clicks on the URL and reaches the site, these parameters are stored in our database. The Web also gives the user the possibility to write a comment or to unsubscribe from the system so that she will no longer receive emails. Maintaining the user subscribed to our system gives us the possibility to perform future experiments.

3.3 Experimental Results

We have performed two separate experiments. In order to generate a baseline, we sent typical spam from a classical spam text in order to measure the success rate, taking into account that spam could have been detected and filtered by the email service, Internet Service Provider, email client in the users computer, or it could have been ignored or deleted by the user. In the second experiment, we focused on personalized spam, in order to prove the click through rate obtained, sending a bigger amount of personalized spam. In both experiments the users were assigned randomly to each of them. The results of each experiment, and explanation thereof will be explained in the next two sections. The comparison of the results is discussed in the last subsection.

3.3.1 First experiment: typical spam

Using multiple email accounts and sending a total of less than one hundred emails per day in order to avoid mail client's restrictions, we sent a typical spam email. The account change is due to a strategy to make things more difficult to spam detection systems. We sent one of those emails where spammers try to draw the receiver's attention to enter a web address. To write this email, we read different emails received in our personal email address and we wrote a similar one. In total, we sent 972 typical spam emails, and results are shown in Table 3.3.

Sent emails	Website visits	Percentage
972	4	0.41%

Table 3.3: Results of the first experiment

As it can be seen, only four users reached our website address. This means that the click-through of the typical spam in our experiment is 0.41%.

3.3.2 Second experiment: personalized spam

In this case, instead of sending typical spam, we sent a personalized emails to 2,889 Facebook users' email addresses. We used the same experiments setup as in the first experiment for the message delivery, and we sent each template from different email accounts. In order to avoid source address blocking, we sent less than one hundred emails per day and account.

As we mentioned previously, we used three different templates in our study. Those templates had a personalized URL to obtain details of each sent email. Note that the website described the experiment, apologizing for the inconvenience caused, and left space for users comments.

Type	Amount	Percentage
Music	1,787	61.85%
Studies	843	29.18%
Company	259	8.97%
Total	2,889	100%

Table 3.4: Number of sent emails

The previous table shows the amount of emails sent, and the their distribution among the generated profile templates. A 'Music' profile-based template was the most commonly used. More than the 60% of the personalized mails encouraged the user to visit an URL regarding their favourite music preferences. Our personalized spam campaign model first checked if the user had music preferences added to her profile. If not, it checked for a past studies profile, and so forth.

As we can see in Table 3.5, 220 users have accessed the website. This is 7.62% of the people that received a personalized email. Also note that 1.38% of people have discharged from the study.

	Amount	Percentage of total shipments
Users who have accessed the website:	220	7.62%
Users who have been unsubscribed:	40	1.38%
Users who have left comments:	11	0.38%

Table 3.5: Website data

Moreover, we break down the answers taking into account the different templates. Table 3.6 shows the website accession from each of the templates sent.

As we can see in the table, most of the users who acceded our website, received a music-related spam message. This can be considered as expected, as music-based templates involve the 61% of the whole campaign. But it is worth highlighting that the 'Company' or work experience-based template got higher click-trough

	Website visits	Percentage of total accesses	Click-through
Music	111	50.45%	6.21%
Studies	81	36.82%	9.61%
Company	28	12.73%	10.81%

Table 3.6: Information according to each template

rate, while the music-based one obtained the lowest one. Otherwise, the musical template had the lowest click-through rate.

3.3.3 Comparison between experiments

	Sent	Answered	Percentage
Typical spam:	972	4	0.41%
Personalized spam:	2,889	220	7.62%

Table 3.7: Comparison between results

Table 3.7 summarizes the response rates obtained while using the different spam types. If we analyze this data, the first interesting information that emerges is that only 4 people have gone through the typical spam. In contrast, 220 other people have come through personalized email. I.e. 0.41 percent compared to 7.62 percent. Authors hypothesize that one reason might be that typical spam can be filtered by most of the spam detection systems. Even so, 0,41% is still many times higher than the rate shown in [84].

3.3.4 Statistical comparison between genders

Finally, in order to see if there are differences between the behaviour of men and women, several statistical comparisons are carried out.

First, we analyzed the gender of the users related to a certain Facebook account to know how many people leave this information publicly available on OSNs. Second, using the emails sent, we are able to extract the information about the amount of women and men that received our email (we didn't send emails to every Facebook account owner, only to a randomly selected ones). Third, we obtain the gender of the users that accessed to our website following the parameterized URL inserted in the emails. And finally, we extract the website visits over the total emails sent per gender. All the mentioned information is presented in the next table.

Gender	FB account	Sent	Website visits	Relative click-through
Man	38.78%	51.96%	56.70%	6.33%
Woman	27.32%	32.47%	26.34%	4.70%
Unknown	33.90%	15.58%	16.96%	6.32%

Table 3.8: Comparison between genders

Table 3.8 shows that the number of men that we could correlate with a Facebook account is significantly higher. Specifically, number of men is 12 percentage points larger than women on our study.

Moreover, if we compare the percentage of sent emails and the visits to our website, it is possible to see that men click more on the URL than women (6.33% vs 4.70%).

Once the general behaviour is shown, information of the sent emails and visits to the website are presented in Table 3.9 divided into the previously defined three different templates.

		Sent		
Gender	Total	Music	Studies	Company
Man	1502	57.39%	33.49%	9.12%
Woman	937	69.37%	23.91%	6.72%
Unknown	450	61.11%	25.78%	13.11%
		Click-through (%)		
Man	8.26%	6.61%	9.74%	13.14%
Woman	6.30%	4.46%	10.27%	11.11%
Unknown	8.22%	9.09%	7.76%	5.08%

Table 3.9: Comparison between genders divided in templates

The difference between the behaviour of men and women is reflected in each template, while the click-through of men is bigger using 'Company' (13.14% vs 11.11%) and 'Music' (6.61% vs 4.46%) templates, using the 'Studies' template the opposite is shown (9.74% vs 10.27%). Presented results show the huge difference in the behaviour of women depending on the template. While the click-through of 'Studies' and 'Company' templates are similar (10.27% and 11.11%), the result obtained with the 'Music' template is 5 points lower (4.46%).

3.4 Countermeasures

After the whole experimentation and results discussion, we consider three ways to avoid spam customization. Two from the OSNs point of view, and the other from the users perspective.

- *Limiting users' public information:* OSNs may limit public information from users. Thus, it might be more difficult to extract information from users. And the attackers can not use this information in their attacks. This is an obvious countermeasure, but authors consider that goes in the opposite position of what OSN owners seek to attract more attention.

- *Changing the code of the website:* At the time of writing this dissertation it is possible to collect information from the source code of the Facebook web page. If they change the website and do not leave the user information in a extractable format (for instance: images), it will be more difficult to obtain information for attackers. An interesting research line could be the use of code randomization that could evade automatic web page scrapping.

- *Raising Awareness:* We must teach people how dangerous it can be to leave personal information publicly. If people minimize their profiles public information will be much more difficult to customize the spam.

- *Content analysis:* New spam detection systems must be developed focussed on personalized spam detection. One way to improve spam filtering should be analysing meticulously the content of the messages applying new natural language processing tools.

3.5 Ethical Considerations

Some actions taken in this chapter are ethically sensitive. For some people, collecting information from the Internet is not ethically correct. But as was discussed in [76, 77] and more recently in [14], the best way to do an experiment is to do as realistically as possible. We defend this mode of action for the following reasons.

First, we must be clear that we work to improve the safety of users, we use users information to protect them in the future. Second, we only use information that users displayed publicly in OSNs. This means that we never attacked any account, password or private area. Third, attackers use this information, if we use the same information and act in the same way, we will defend users better.

Finally, we have consulted to the leadership of our university and they have given us the approval. For this, we proposed our intentions to the general direction

of the university before the experiments took place (spam campaigns), where we showed them the ethical considerations for conducting the study. We also explained them the procedure we had designed to collect personal data and the way we had thought to send emails. Once the R&D Manager gave us the approval, we started with the experiment phase.

3.6 Conclusions of the Chapter

This chapter makes clear the issue that could exist if spam campaign creators turn their spam templates into a personalized text based on user characteristics, interests, and motivating subjects. Attackers have millions of email addresses stored. We have demonstrated that a 19% of the collected email addresses have a corresponding Facebook account associated to it. Moreover, basic public information can be extracted from those users, which is sufficient to create personalized email subject and bodies. These emails can have a click-through rate higher than a 7.62%, being this more than 1,000 times higher than typical spam campaign rate s as shown in [84]. It is obvious that in parallel to the research of new techniques for spam detection inside OSNs, it is necessary to perform research beyond the state of art of classic spam filtering, taking into account the possibility of personalized spam campaign success.

Regarding the behaviour of OSN users analyzed, we found that most of the Facebook users choose their favourite music band and leave it public. We could also see that 30% of users who have some data in their Facebook profile, have at least one company with which they have been connected.

Another interesting fact is that there are more men than women associated with their email to Facebook. The number of men is 12 percentage points higher than the number of women. There is also a difference between genders in the click-through rate, while women react in a rate of a 4.70%, men do it in a 6.33%. Consciousness differences and gender psychological reasons might arise to explain this fact.

The main conclusion to be drawn is that we can develop advanced techniques for generating personalized mail that circumvent current spam detection systems. Clear examples of this are the results shown in the results section. In the first experiment, we can see that only the 0.41% of users have bitten the bait. Whereas in the second 7.62% of the users have entered to the project website. The second result rate is more than 18 times higher than the first one.

We can see that it is not a large number of people, but as a steady stream of visitors, which means that personalized emails reach their destination. Then, once the message is on the user's email inbox, it depends on each person's behaviour to click on the link that is sent in the mail.

CHAPTER 4

INFLUENCE OF SENTIMENT ANALYSIS IN SPAM FILTERING

This chapter provides means to validate the assumption that being a spam message a commercial communication, the semantics of its content should be shaped with a positive meaning. Thus, the main objective of this chapter is to analyze if the polarity of the message is a useful feature for spam classification. It also aims to validate the hypothesis that polarity feature can improve the results of the typical spam filtering techniques.

To do that, after a brief analysis of the related work, our method is presented where on the one hand, we apply the most effective spam classification filters to a original dataset, whereby we obtain the algorithms that better classify the content into spam and ham classes. On the other hand, we analyze different settings of two sentiment classifiers: an API for diving into common NLP tasks and others developed by us. Once we select the classifiers and settings that obtained the best results in the analysis, we determine the polarity of the texts in the original dataset, and we create new datasets adding the polarity feature to each text. Then a descriptive analysis of the new datasets is carried out. Finally we apply the spam filtering classifiers that obtained the best results in the original dataset to the new ones.

In the next section, three different scenarios are taken into account in order to provide solid means to validate our hypothesis.

The first validation experiment is carried out focusing on email spam messages which is a very common problem in our society, as it is explained in Chapter 2.

In the second part, a similar work-flow is followed. In this case, instead of using

email spam data, we use SMS messages. The influence of polarity in short instant messages spam filtering is analyzed.

Finally, we validate the impact of using sentiment analysis techniques for the detection of social spam, using a spam dataset from a very popular OSN.

The main contribution of this chapter is that we improve spam filtering rates using the polarity.

4.1 Relevant Proposals

While most researchers are working on opinion spam detection using NLP and/or text mining techniques, we focus on the use of NLP and text mining techniques in conjunction with Sentiment Analysis (SA) to improve the detection of spam emails.

In [98], SA or opinion mining is defined as the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes. In SA NLP, text analysis and computational linguistics are used to identify and extract subjective information in source material. As explained in [119], the area of SA has had a huge burst of research activity during these last years, but there has been a continued interest for a while. Currently there are several research topics on opinion mining and the most important ones are explained in [98]. Among those topics we identified the document sentiment classification as a possible option for spam filtering.

The main objective of this area is classifying the positive or negative character of a document [119]. In order to classify such sentiment, some researchers use supervised learning techniques, where three classes are previously defined (positive, negative and neutral) [120]. Some other authors propose the use of unsupervised learning. In unsupervised learning techniques, opinion words or phrases are the dominating indicators for sentiment classification [154].

There are several tools developed during the last years focused on NLP and sentiment analysis. One of the most used for sentiment analysis is known as SentiWordNet. The first version was presented in [42] and a improved one was released by Baccianella et al. [12] some years later. It is an enhanced lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. As they explained in the paper SentiWordNet is the result of the automatic annotation of all the synsets of WordNet according to the notions of positivity, negativity, and neutrality. For instance, the authors in [116] used SentiWordNet for sentiment classification of reviews obtaining an accuracy of 65.85 % using term counting method.

4.2 Proposed Method

This method is divided in three different phases.

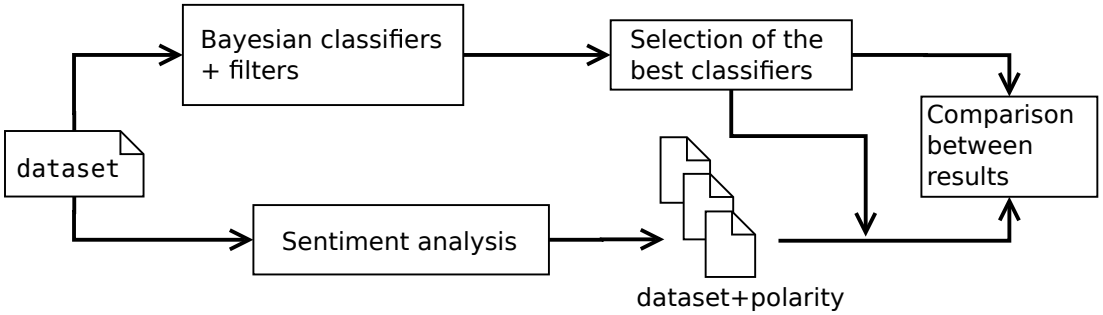

Figure 4.1: Improving spam filtering using sentiment analysis

As Figure 4.2 shows, first, we apply several spam filtering models with different settings to a certain dataset. Thus, we identify the best classifiers and the best settings to filter spam messages.

As our objective is to improve the best classifiers, in the second phase we work to obtain better results than previously mentioned filters using the polarity of the text. For that, first of all we need to determine the polarity of each text. To do so, we developed our own sentiment classifier, and also used a publicly available API for NLP tasks known as TextBlob[1]. Comparing different settings of each classifier we selected the best ones, which were applied to the dataset used in the previous phase. Using the polarity of each message as new attribute, we carry out a descriptive analysis of these datasets.

Finally, the best spam classifiers were applied to the new datasets and made a comparison of the results. In order to validate the obtained the results, the experiment is repeated using another dataset.

All the experiments are carried out using the 10-fold cross-validation technique and the results are analyzed in terms of the number of the false positive and accuracy, being the accuracy the percentage of testing set examples correctly classified by the classifier.

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

4.2.1 Bayesian spam filtering

Those filters, which are based on Bayes' Theorem, use Bayes logic to evaluate the header and content of an incoming e-mail message and determine the probability that it constitutes spam.

[1]http://textblob.readthedocs.org

The main objective is to identify the best spam filtering classifiers and the best settings. We apply different combinations of classifiers, filters and settings to compare the results and to select the best ones.

As explained in Section 4.1, Bayesian classifiers are considered as the best techniques to detect and to filter spam. In this chapter, the next Bayesian classifiers have been used:

- Large-scale Bayesian logistic regression for text categorization.
- Complement class Naive Bayes classifier.
- Discriminative parameter learning for Bayesian networks.
- Updateable multi-nominal Naive Bayes classifier.
- Naive Bayes.
- Multi-nominal Naive Bayes classifier.
- Naive Bayes Updateable.

Following a text mining process, a set of different filters have been applied to the text. Next, we detail the settings that have been used:

- A filter to convert a string to feature vector containing the words. We use the next options:
  - Words are converted to lower case.
  - A number of words to keep is defined.
  - The maximum number of words and the minimum term frequency is not enforced on a per-class basis but based on the documents in all the classes.
  - Two type of tokenizers are used:
    - One that splits the text removing the special characters.
    - And the other that removes the characters and to split a string into an n-gram with min and max grams.
  - To obtain roots of the words a stemmer based on the Lovins stemmer is used.
  - Weights:
    - IDFTransform (Inverse Document Frequency (IDF) Transformation) False, TFTransform (Term Frequency score (TF) Transformation) False, outputWordCounts (counts the words occurrences) False

    - IDFTransform False, TFTransform False, outputWordCounts True
    - IDFTransform True, TFTransform False, outputWordCounts True

- Attribute Selection: a ranker to evaluate the worth of an attribute by measuring the information gain with respect to the class is used.

At the end of this phase, the best ten settings and classifiers for spam classification have been identified. To do this selection we have use the accuracy of the classifiers.

4.2.2 Sentiment analysis

The objective of this phase is to carry out a sentiment classification of the dataset, in order to later add the polarity of each text as a new feature for spam detection. After that, the influence of the polarity in spam filtering is analyzed.

First a sentiment classifier is needed, so in this task two different options have been considered: to develop our own classifier or to use an existing one. In order to obtain the best possible results, both options have been considered.

Sentiment classifier based on SentiWordNet dictionary. In order to design and implement a classifier, sentiment dictionaries become useful tools, so the commonly used SentiWordNet has been chosen in this case. As shown in researchers have obtained up to a 65% of accuracy using this dictionary.

SentiWordNet is a dictionary that returns to the user the polarity of a certain word depending on its grammatical properties. Using this tool, the average polarity of the email messages have been calculated.

Five sentiment classifiers have been developed with different settings. On the one hand: *Adjective*, *Adverb*, *Verb* and *Noun*. In each classifier every word was considered to be a certain part of speech (depending on the name of the classifier), so we have obtained the polarity of those words that have that grammatical property. For instance: in the *Adjective* classifier every word was considered to be an adjective, so we have obtained the polarity of those words that can be considered as adjectives. And on the other hand, *AllPosition* classifier, which considers every part of speech per each word.

TextBlob classifier. With the objective of comparing different results, TextBlob has been used because it provides a simple API for diving into common NLP tasks. Specifically, giving a string the sentiment analyzer function returns a float value within the range [-1.0,1.0] for the polarity.

To improve the effectiveness of those classifiers we change settings and select different thresholds (-0.05, -0.1, 0, 0.1, 0.05). The threshold means the point were

we consider the polarity score positive or negative, and we use it in the name of the classifier to differentiate from each other.

Descriptive experiments

In this step, several experiment have been carried out to see how sentiment analysis can affect in spam filtering.

First of all, the selected three classifiers have been applied to the dataset which is explained in the following section. This step offers an idea about the distribution of the email messages in terms of polarity. The number and the percentage of the positive and negative messages has been obtained. Moreover, this information has been used to created one file per each selected classifier, in which the polarity of each message has been added.

Then, we generate a ranking of the most important attributes based on the information gain criteria, and also by analyzing the features that better divide a *J48* classification tree node. Doing that, we preliminarily analyze how the polarity affects in terms of spam filtering.

Predictive experiments

During this task the classifiers that obtained the best results in the spam filtering experiments has been applied to the different datasets files. Those files have been created during the descriptive experiments and it consists in a certain spam dataset with the polarity of each message. So, at the end of the experiment the accuracy of the best 10 classifiers applied to the sentimentally classified messages have been obtained. And finally, all the results are compared.

4.3 Validation of the Proposed Method

In this Section the validation of the proposed method is presented, taking into account the results of three different scenarios.

4.3.1 Email Spam

The objective of this scenario is to analyzed the influence of sentiment analysis in email spam filtering.

To do that, the procedure explained in Section 4.2 is followed.

Email datasets used in sentiment analysis

To carry out this test those datasets are used:

- *Movie Reviews*[2]: This dataset collects movie-review documents tagged in terms of polarity (positive or negative) or subjectivity rating. Also sentences are tagged with respect to their status or polarity. Among all these options the *polarity dataset v2.0* is used in this task, which is composed of 1,000 positive and 1,000 negative processed reviews introduced in [118]. This dataset is used to evaluate the effectiveness of each sentiment classifier during.

- *CSDMC 2010 Spam Corpus*[3]: composed of 2,949 legitimate email messages and 1,378 spam messages. This dataset is used as to carry out the original experiments.

- *TREC 2007 Public Corpus*[4]: This corpus contains 75,419 email messages: 25,220 ham (legitimate) and 50,199 spam emails. And we use it to repeat the experiment and to validate the results obtained using the previous dataset. In order to work carry out the experiments using similar datasets in terms of email number, 4,000 emails are selected randomly (3,000 ham and 1,000 spam).

Sentiment analysis of emails

Once the classifiers have been defined in Section 4.2, we improve the efficiency of those classifiers by changing settings and selection thresholds. For this work, a previously tagged dataset is mandatory, and in this case *Movie Reviews* dataset is used. The objective is to obtain the best accuracy classifying those reviews to find the most efficient settings and thresholds.

In Table 4.1, a comparison between the best settings and thresholds is shown. The next criteria is used to define each classifier:

- *TextBlob* means that a classifier based on Textblob library has been used.

- Some names are followed by a number. This number represent the used threshold for polarity classification. For instance, 0.1 means that every message with score higher than 0.1 has been considered to be positive, and those message with score lower than 0.1 negatives.

- "All without verbs" means that all part of speech but verb has been taken into account during the score calculation.

Name	TP	TN	FP	FN	Accuracy
TextBlob 0.1	719	773	227	281	0.75
TextBlob 0.05	901	467	533	99	0.68
Adjectives	775	499	501	225	0.64
All without verbs	798	460	540	202	0.63
AllPositions	849	370	630	151	0.61
Nouns	723	483	517	277	0.60
TextBlob 0	971	229	771	29	0.60

Table 4.1: Comparison between classifiers

Using this information the best three classifiers are selected. To decide which ones can be considered as the best classifiers, the *Accuracy* measure is used.

Experimental results

In this section, the results obtained during the previously explained experiments are shown. To carry out those experiment the *CSDMC2010* is used.

Bayesian spam filtering experiment. First of all, Bayesian classifiers with different settings have been applied to the *CSDMC2010* dataset. In total, 392 different combinations, defined in Section 4.2, are analyzed. In Table 4.2 the best 10 classifiers in terms of accuracy are shown.

#	Spam classifier	TP	TN	FP	FN	Acc
1	BLR.i.t.c.stwv.go.wtok	1,355	2,936	13	24	99.15
2	DMNB.c.stwv.go.wtok	1,362	2,928	21	17	99.12
3	DMNB.i.c.stwv.go.wtok	1,362	2,928	21	17	99.12
4	DMNB.i.t.c.stwv.go.wtok	1,362	2,928	21	17	99.12
5	DMNB.stwv.go.wtok	1,362	2,928	21	17	99.12
6	DMNB.c.stwv.go.stemmer	1,360	2,927	22	19	99.05
7	DMNB.i.c.stwv.go.stemmer	1,360	2,927	22	19	99.05
8	DMNB.i.t.c.stwv.go.stemmer	1,360	2,927	22	19	99.05
9	DMNB.stwv.go.stemmer	1,360	2,927	22	19	99.05
10	BLR.i.t.c.stwv.go.ngtok.stemmer.igain	1,351	2,935	14	28	99.03

Table 4.2: Top10 Bayesian classifiers

[2]http://www.cs.cornell.edu/People/pabo/movie-review-data/
[3]http://www.csmining.org/index.php/spam-email-datasets-.html
[4]http://plg.uwaterloo.ca/ gvcormac/treccorpus07/

In this study, the objective is to improve the accuracies of the Bayesian classifier. So, we focus only on these 10 classifiers in the following steps, instead of focus on all combinations used previously.

The used nomenclatures are explained in Table 1.

Sentiment analysis

Descriptive experiments. During the data exploration part, the following results are presented.

Firstly, a sentiment analysis of the dataset has been done. The polarity of each email is identified, this polarity is added to the dataset and statistics of the number of positive and negative spam or legitimate emails are extracted as it is shown in the Table 4.3. As we showed that an important number of messages obtained score equal to 0 using *Adjective* classifier, *Adjective Plus* classifier is added in this point. It classified those emails like positive messages. So at the end of this step four different dataset are created, one per each classifier.

		Adj		**Adjplus**		**Tb 005**		**Tb 01**	
	Total	*P*	*N*	*P*	*N*	*P*	*N*	*P*	*N*
spam	*1,378*	913	433	945	433	1,044	332	848	516
ham	*2,949*	1,103	1,831	1,118	1,831	1,934	1,009	1,419	1,514
Percentages (%)									
spam	100	66	31	68	31	76	24	62	37
ham	100	37	62	37	62	66	34	48	51

Table 4.3: Sentiment analysis of emails

Analysing the data in Table 4.3, it is possible to see that spam messages are more positive than non-spam or ham messages.

While this experiments gives good results, the results obtained in the rankings and in the trees were not such good. We observed that polarity appears like a decisive attribute but not like a top one. And different results have been obtained depending on the used sentiment classifier. The best results have been obtained by the dataset analized by the *Adjective* classifier. The polarity is ranked in the position 130, and is considered a bit decisive attribute in *J48* decision tree. *Adjective Plus* classifier obtains similar but worse results. And significantly worse results have been obtained by TextBlob-based classifiers.

Predictive experiments and comparison. Once known that polarity can affect in spam filtering, an experiment to demonstrate the real influence is carried out. The best classifiers that appears in Table 4.2 are applied to the four new datasets.

In Tables 4.4 and 4.5 the results are displayed. In both tables the results obtained during the Bayesian filtering are shown for a proper comparison between the results.

On the one hand, in Table 4.4 the original results are compared with the ones obtained using the dataset tagged by our own developed classifier.

	Sentiment analyzer								
	None			*Adjective*			*Adjective+*		
#	FP	FN	Acc	FP	FN	Acc	FP	FN	Acc
1	13	24	99.15	14	22	**99.17**	14	23	99.15
2	21	17	99.12	24	15	99.10	24	15	99.10
3	21	17	99.12	24	15	99.10	24	15	99.10
4	21	17	99.12	24	15	99.10	24	15	99.10
5	21	17	99.12	24	15	99.10	24	15	99.10
6	22	19	99.05	21	17	**99.12**	22	16	**99.12**
7	22	19	99.05	21	17	**99.12**	22	16	**99.12**
8	22	19	99.05	21	17	**99.12**	22	16	**99.12**
9	22	19	99.05	21	17	**99.12**	22	16	**99.12**
10	14	28	99.03	14	24	**99.12**	15	23	**99.12**

Table 4.4: Comparing original results with the results obtained using own polarity classifiers

As we can see in those first results, the *Adjective* sentiment classifier is able to improve the best accuracy of Bayesian algorithms.

On the other hand, in Table 4.5, the original results are compared with the results obtained applying the filtering classifiers to the dataset tagged by TextBlob-based classifiers.

If we analyze this data, we can realize that polarity helps to improve the accuracy in most cases, and also that the best result obtained using Bayesian spam filtering is improved. While without polarity the best result is 99.1451%, using the polarity feature we reached the rate of 99.2144%.

Focusing on the results of the *TextBlob01* sentiment classifier, we see that in eight out of ten cases the accuracy is better than in the original result. And in case number 9 the same accuracy is obtained.

	Sentiment analyzer								
	None			*TextBlob 005*			*TextBlob 01*		
#	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	13	24	99.15	13	25	99.12	14	24	99.12
2	21	17	99.12	24	15	99.10	22	12	**99.21**
3	21	17	99.12	24	15	99.10	22	12	**99.21**
4	21	17	99.12	24	15	99.10	22	12	**99.21**
5	21	17	99.12	24	15	99.10	22	12	**99.21**
6	22	19	99.05	21	15	**99.17**	22	15	**99.15**
7	22	19	99.05	21	15	**99.17**	22	15	**99.15**
8	22	19	99.05	21	15	**99.17**	22	15	**99.15**
9	22	19	99.05	21	15	**99.17**	22	15	**99.15**
10	14	28	99.03	14	24	**99.12**	14	28	99.03

Table 4.5: Comparing original results with the results obtained using TextBlob polarity classifiers

Second dataset. In order to validate results from the first part, the experiment is repeated using another dataset. In this case, we use the previously presented TREC2007 dataset. And the same ten classifiers that obtained the best results with the previous dataset are applied to TREC2007. The obtained results are shown in Table 4.6.

#	**Spam classifier**	**TP**	**TN**	**FP**	**FN**	**Acc**
1	BLR.i.t.c.stwv.go.ngtok.stemmer.igain	976	2,983	17	24	98.98
2	DMNB.c.stwv.go.stemmer	979	2,979	21	21	98.95
3	DMNB.i.c.stwv.go.stemmer	979	2,979	21	21	98.95
4	DMNB.i.t.c.stwv.go.stemmer	979	2,979	21	21	98.95
5	DMNB.stwv.go.stemmer	979	2,979	21	21	98.95
6	DMNB.c.stwv.go.wtok	977	2,979	21	23	98.90
7	DMNB.i.c.stwv.go.wtok	977	2,979	21	23	98.90
8	DMNB.i.t.c.stwv.go.wtok	977	2,979	21	23	98.90
9	DMNB.stwv.go.wtok	977	2,979	21	23	98.90
10	BLR.i.t.c.stwv.go.wtok	972	2,978	22	28	98.75

Table 4.6: Results of the best 10 classifiers applied to the validation dataset

Once the original results are obtained, we carry out a sentiment analysis of the TREC2007 dataset and we add the polarity feature to it, creating four new datasets (one per each sentiment analyzer). Finally, the best ten classifiers are applied to the new datasets and the obtained results are presented.

	Sentiment analyzer								
	None			*Adjective*			*Adjective+*		
#	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	17	24	98.98	17	21	99.05	17	21	99.05
2	21	21	98.95	20	18	99.05	20	18	99.05
3	21	21	98.95	20	18	99.05	20	18	99.05
4	21	21	98.95	20	18	99.05	20	18	99.05
5	21	21	98.95	20	18	99.05	20	18	99.05
6	21	23	98.90	20	23	98.93	20	23	98.93
7	21	23	98.90	20	23	98.93	20	23	98.93
8	21	23	98.90	20	23	98.93	20	23	98.93
9	21	23	98.90	20	23	98.93	20	23	98.93
10	22	28	98.75	22	31	98.68	22	31	98.68

Table 4.7: Comparing original results with the results obtained using own polarity classifiers

	Sentiment analyzer								
	None			*TextBlob005*			*TextBlob01*		
#	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	17	24	98.98	16	20	**99.10**	17	21	99.05
2	21	21	98.95	21	21	98.95	21	21	98.95
3	21	21	98.95	21	21	98.95	21	21	98.95
4	21	21	98.95	21	21	98.95	21	21	98.95
5	21	21	98.95	21	21	98.95	21	21	98.95
6	21	23	98.90	19	23	98.95	21	23	98.90
7	21	23	98.90	19	23	98.95	21	23	98.90
8	21	23	98.90	19	23	98.95	21	23	98.90
9	21	23	98.90	19	23	98.95	21	23	98.90
10	22	28	98.75	22	34	98.60	22	34	98.60

Table 4.8: Comparing original results with the results obtained using TextBlob polarity classifiers

Tables 4.7 and 4.8 show that the results are improved in almost all the cases in terms of *accuracy*, and also that the number of false positive is reduced. Moreover, the original best result is also improved from 98.98% to 99.10%, reducing the number of false positives.

4.3.2 SMS Spam

To analyze the influence of the sentiment in SMS spam filtering, this part has been carried out following the procedure showed in Section 4.2. But it in this case two experiments are carried out applying different classifiers to the datasets in order to demonstrate the influence of the polarity over the SMS spam filtering.

SMS datasets used in sentiment analysis

During this study two different publicly available datasets are used:

- *SemEval-2013*[5]: Introduced in [110]. This dataset contains labelled (positive, negative and neutral) mobile phone messages, and we use it to evaluate the effectiveness of each sentiment classifier during the first phase. Specifically we use positive (492 SMS) and negative (394 SMS) messages.

- *SMS Spam Collection v.1*[6] (called SMSSpam in this dissertation): Published in [7], it is composed of 5,574 English, real and non-enconded messages, tagged according 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*[7] (called BritishSMS in this dissertation): Introduced in [113]. 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.

Sentiment analysis of SMS messages

The main objective of this part is to add the polarity of each message to the original dataset *SMSSpam* in order to carry out the experiments.

As in the previous experiments, in order to analyze different sentiment classifiers, we follow the same process than in 4.3.1.

[5]https://www.cs.york.ac.uk/semeval-2013/task2/
[6]http://www.dt.fee.unicamp.br/ tiago/smsspamcollection/
[7]https://goo.gl/UUgl4X

Once the classifiers have been defined in Section 4.2, a tagged (in terms of polarity) publicly available dataset is required to evaluate the effectiveness of the classifiers. Taking into account that SMS messages are the final objective, we decided to choose a dataset composed by SMS messages in order to obtain more reliable results: *SemEval-2013* dataset.

We apply different sentiment classifiers to the dataset, and we analyze the accuracy of correctly classified messages. In Table 4.9 a comparison between different classifiers and thresholds is shown.

Classifier	**Accuracy**	**Classifier**	**Accuracy**
TextBlob 0.05	0.78	Adjectives	0.66
TextBlob 0.1	0.76	Nouns	0.58
TextBlob -0.05	0.73	TextBlob 0	0.56
AllPositions	0.72	Adverb	0.53
TextBlob -0.1	0.71	Verb	0.52

Table 4.9: Comparison in terms of accuracy between the best classifiers

Based on the accuracies offered by the table, 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.

SMS Spam filtering

To analyze the influence of the polarity over the filtering of SMS messages, the first step is to select 10 representative classifiers and some of the best filter settings for natural language processing. To do that, the results presented in Section 4.3.1 are taken into account. Consequently the best five classifiers from the mentioned section are used in this study. Also the best three settings of the filters are chosen. Moreover, we added more classifiers to the list based on other research studies such as [91]. Final list: Large-scale Bayesian logistic regression for text categorization, discriminative parameter learning for Bayesian networks, complement class Naive Bayes classifier, multi-nominal Naive Bayes classifier, updateable multi-nominal Naive Bayes classifier, decision tree (C4.5), random tree, forest of random trees, Support Vector Machine (SMO) and adaptive boosting meta-algorithm with Naive Bayes.

The next step is to apply those classifiers, combined with the best three filters and settings, to the datasets (with and without polarity) and compare the results.

This first step provides the best classifier for text messages, so in the following phase, the best six classifiers are picked. And the second experiment is carried out applying those classifiers with different combination of filters and settings (56 combinations per classifier) to the datasets. The objective of the combination of these filters and settings is to follow a text mining process in order to compare results and identified the best ones, and those are some of its main details:

- A filter to convert a string to feature vector containing the words. We combine different options: words are converted to lower case, special characters (.,;:$&%=_@()?!+-#[]) are removed using tokenizers; n-gram with min and max grams are created; roots of the words are obtained using a stemmer, etc.

- Attribute Selection: a ranker to evaluate the worth of an attribute by measuring the information gain with respect to the class (spam/ham) is used.

Using those combinations, we identify the best ten settings and classifiers for SMS spam classification, and those are applied to the dataset with polarity to compare the results.

Experimental results

In this Section the results obtained during the previously explained experiments are shown. To carry out those experiment the dataset called *SMS Spam Collection v.1* is used.

Descriptive experiment. Once the dataset is selected, we perform a descriptive experiment of the dataset. The objective of this step is to analyze the polarity of the messages applying our previously selected sentiment classifiers. This is the point where the polarity extracted during the analysis is inserted in the dataset, creating three new datasets (one per each classifier) and where statistics about the polarity are calculated.

In Table 4.10, the results of the experiment are presented (*Tb 005* means *TextBlob 0.05*, *Tb 01* means *TextBlob 0.1* and *Tb -005* means *TextBlob -0.05*).

According to the classifiers, it is possible to see, specially in the first two sentiment classifiers, 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.

Two experiments to see the real influence are carried out.

	Number of messages						Percentage (%)					
	Tb 005		**Tb 01**		**Tb -005**		**Tb 005**		**Tb 01**		**Tb -005**	
	P	N	P	N	P	N	P	N	P	N	P	N
spam	430	317	408	339	688	53	58	42	55	45	92	7
ham	1,960	2,867	1,859	2,968	4,121	687	41	59	39	61	85	14

Table 4.10: Sentiment analysis of SMS messages

Finding the best SMS spam filtering classifiers. This experiment aims to identify the best SMS spam filtering classifiers in order to use them in the next experiment with more filters and settings. As it is mentioned in Section 4, we choose 10 classifiers and the following filter combinations per each one. Those filters are used to obtain the results presented in the Table 4.11. The explanation is based on the results obtained in Section 4.3.1.

1. *stwv.go.wtok*: the best result.

2. *i.t.c.stwv.go.ngtok.stemmer.igain*: into the best two algorithms these settings obtained the best result in one, and the second result in the other.

3. *i.t.c.stwv.go.wtok*: appeared in the top ten results and is was the first with n-grams and information gain filter.

The nomenclature used in this list and in the following tables is explained in Table 1.

In Table 4.11, the results of the ten classifiers with the three listed filters are presented. The number in the name represents the type of the filter used.

Analyzing the table we can see that applying classifiers to the original dataset the best one in terms of accuracy is SMO with the third settings. In this case, the polarity does not help to improve the result. But applying the Discriminative Parameter Learning for Bayesian Network (DMNBtext) classifier and the first filter to the dataset created using the sentiment analyzer *TextBlob01* the top result is improved. Specifically, an accuracy of 98.76% is obtained. In other two cases the top result is also improved.

Another important information shown in Table 4.11 is that although it is not the best result in terms of accuracy, there is case that must be highlighted. This case is the Bayesian Logistic Regression with the second filter applied to the dataset *TextBlob01*, which obtained an accuracy of 98.67% and 0 false positives.

	Sentiment analizer							
	None		*Tb 005*		*Tb 01*		*Tb -005*	
Spam classifier	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**
SMO.3	3	98.73	5	98.71	4	98.67	3	98.73
NBM.3	12	98.69	12	98.62	12	98.69	8	98.69
NBMU.3	12	98.69	12	98.62	12	98.69	8	98.69
BLR.3	5	98.64	6	98.60	5	98.67	4	98.64
DMNB.1	10	98.62	7	**98.74**	7	**98.76**	6	98.69
BLR.2	2	98.60	3	98.49	**0**	**98.67**	**0**	**98.58**
NBM.1	23	98.53	23	98.51	22	98.53	16	98.64
SMO.2	4	98.53	4	98.58	5	98.58	4	98.53
NBMU.2	36	98.51	36	98.49	35	98.51	26	98.67
NBMU.1	29	98.49	25	98.56	26	98.55	16	98.73
CNB.1	31	98.44	32	98.40	32	98.39	18	98.64
NBM.2	52	98.37	52	98.33	52	98.31	46	98.48
DMNB.2	4	98.28	3	98.31	3	98.37	2	98.31
DMNB.3	4	98.28	3	98.31	3	98.37	2	98.31
CNB.2	64	98.19	64	98.15	62	98.19	58	98.30
CNB.3	56	98.17	48	98.30	48	98.26	19	**98.74**
BLR.1	1	97.45	0	96.41	0	96.59	0	96.18
SMO.1	0	97.45	0	97.54	0	97.56	0	97.45
J48.3	54	97.02	58	96.68	56	96.72	54	96.97
J48.2	58	96.90	62	96.56	62	96.54	58	96.86
J48.1	42	96.86	43	96.90	43	96.90	42	96.86
RF.2	0	96.38	0	95.82	2	96.05	0	96.39
RF.3	0	96.21	0	96.27	1	95.91	0	96.29
RT.1	25	95.60	18	95.59	24	95.71	17	95.95
RF.1	0	95.19	0	94.76	0	94.94	0	95.03
RT.3	84	95.16	79	95.43	92	94.90	95	95.25
RT.2	88	95.07	73	95.35	79	95.28	93	95.32
ABM.2	167	91.44	166	91.46	166	91.46	167	91.44
ABM.3	167	91.44	166	91.46	166	91.46	167	91.44
ABM.1	188	91.32	139	91.59	139	91.59	188	91.32

Table 4.11: Comparison between results

#	Spam classifier	TP	TN	FP	FN	Acc
1	NBMU.i.c.stwv.go.ngtok	1,355	2,936	13	24	99.15
2	NBMU.i.t.c.stwv.go.ngtok	1,362	2,928	21	17	99.12
3	NBM.i.t.c.stwv.go.ngtok	1,362	2,928	21	17	99.12
4	NBMU.i.t.c.stwv.go.ngtok.stemmer	1,362	2,928	21	17	99.12
5	NBM.c.stwv.go.wtok	1,362	2,928	21	17	99.12
6	NBM.i.t.c.stwv.go.ngtok.stemmer	1,360	2,927	22	19	99.05
7	NBMU.c.stwv.go.wtok	1,360	2,927	22	19	99.05
8	CNB.i.t.c.stwv.go.ngtok.stemmer	1,360	2,927	22	19	99.05
9	NBM.i.c.stwv.go.ngtok	1,360	2,927	22	19	99.05
10	NBM.i.c.stwv.go.ngtok.stemmer	1,351	2,935	14	28	99.03

Table 4.12: Top10 Bayesian classifiers

SMS spam filtering with polarity score. The second experiment is based on the results obtained in the first one. While the previous aims to search the best algorithms, this one aims to explore most of the possible filter combinations with the best classifiers.

On this way, we identify the best 6 classifiers in Table 4.11 and combined each one with 56 different filter settings. We analyze the achieved results and we get the classifiers that obtained the best ten results in terms of accuracy as Table 4.12 shows.

The next step is to apply those classifiers to the new datasets that we created using the sentiment classifiers. Those results are shown in Table 4.13.

	Sentiment analyzer											
	None			*Tb 005*			*Tb 01*			*Tb -005*		
#	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	28	36	98.85	36	35	98.73	36	35	98.73	35	35	98.74
2	27	39	98.82	17	61	98.60	16	56	98.71	8	61	98.76
3	32	36	98.78	37	33	98.74	37	33	98.74	33	35	98.78
4	23	45	98.78	36	36	98.71	36	36	98.71	34	36	98.74
5	13	56	98.76	33	35	98.78	32	35	98.80	28	36	98.85
6	34	35	98.76	34	36	98.74	33	37	98.74	32	37	98.76
7	13	56	98.76	17	61	98.60	16	56	98.71	8	61	98.76
8	37	34	98.73	28	36	98.85	28	36	98.85	27	39	98.82
9	37	34	98.73	26	38	98.85	25	38	98.87	22	39	**98.91**
10	36	35	98.73	23	44	98.80	22	44	98.82	19	47	98.82

Table 4.13: Comparing original results with the results obtained using sentiment analyzers

Table 4.13 that a higher accuracy than in the previous experiment is obtained applying new settings of the filters to the original SMS dataset.

The data 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* 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 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 false positives in the other.

Second dataset. As mentioned previously, in order to validate the obtained results during the previous experiment, we decided to repeat the test but using a different dataset. In this case, we use the publicly available BritishSMS dataset.

First of all, the best ten spam classifiers identified in the previous experiment are applied to the dataset, obtaining the following results:

#	**Spam classifier**	**TP**	**TN**	**FP**	**FN**	**Acc**
1	NBM.i.c.stwv.go.ngtok	1,355	2,936	13	24	99.15
2	NBM.i.t.c.stwv.go.ngtok	1,362	2,928	21	17	99.12
3	NBMU.i.c.stwv.go.ngtok	1,362	2,928	21	17	99.12
4	NBMU.i.t.c.stwv.go.ngtok	1,362	2,928	21	17	99.12
5	NBMU.i.t.c.stwv.go.ngtok.stemmer	1,362	2,928	21	17	99.12
6	NBM.i.c.stwv.go.ngtok.stemmer	1,360	2,927	22	19	99.05
7	CNB.i.t.c.stwv.go.ngtok.stemmer	1,360	2,927	22	19	99.05
8	NBM.i.t.c.stwv.go.ngtok.stemmer	1,360	2,927	22	19	99.05
9	NBM.c.stwv.go.wtok	1,360	2,927	22	19	99.05
10	NBMU.c.stwv.go.wtok	1,351	2,935	14	28	99.03

Table 4.14: Results of the best 10 classifiers using the BritishSMS dataset

In the next step, we carry out a sentiment analysis of the BritishSMS dataset, using the same three sentiment analyzers used in the previous experiment, and we add the polarity feature to the original dataset. Doing that, three new tagged dataset are created.

As in the previous experiment, the best ten classifiers are applied to the three new datasets in order to compare the results with the results presented in Table 4.14.

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

	Sentiment analyzer											
	None			*Tb 005*			*Tb 01*			*Tb -005*		
#	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	5	17	97.49	5	17	97.49	5	17	97.49	5	17	97.49
2	5	18	97.37	5	18	97.37	5	18	97.37	5	18	97.37
3	7	16	97.37	7	16	97.37	7	16	97.37	6	16	97.49
4	6	17	97.37	6	17	97.37	6	17	97.37	6	17	97.37
5	9	16	97.14	9	16	97.14	9	16	97.14	9	17	97.03
6	8	18	97.03	8	18	97.03	8	18	97.03	8	18	97.03
7	8	19	96.91	8	19	96.91	8	19	96.91	7	19	97.03
8	8	19	96.91	8	19	96.91	8	19	96.91	7	19	97.03
9	9	23	96.34	9	23	96.34	9	23	96.34	6	24	96.57
10	9	23	96.34	9	23	96.34	9	23	96.34	6	24	96.57

Table 4.15: Comparing original results with the results obtained using sentiment analyzers

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.

4.3.3 Social Media Spam

Once the different experiments using email spam datasets and SMS spam datasets are finished, we aim to validate the message polarity-driven spam detection in OSNs.

First, several Bayesian classifiers are applied to the dataset in order to identify the best classifiers. Second, a descriptive analysis of the dataset is done to see the difference between spam and ham texts. Finally, new datasets are created adding the polarity feature to the original dataset, and the best ten classifiers are applied to them to compare the results with and without polarity.

OSN spam datasets used in sentiment analysis

To carry out this experiment a publicly available dataset is used:

Youtube Comments Dataset [8]: Presented in [115]. This dataset contains multilingual 6,431,471 comments from a popular social media website, Youtube[9]. Among all the comments, 481,334 are marked as spam.

[8] http://mlg.ucd.ie/yt/
[9] www.youtube.com

In order to use similar number of texts to the experiments presented in Section 4.3.1 we created a new subset composed of 1,000 spam and 3,000 ham comments. Those texts have been selected randomly and only taking into account comments written in English.

Experimental results

The experimental phase is divided in two main parts: on the one hand, the descriptive experiment of the dataset is shown, and on the other hand, the predictive experiments and the comparison between the results are carried out.

Descriptive experiment. Like in the previous sections, different sentiment classifiers are needed to perform this experiment. In this case, taking into account the similarities between the Youtube comments and the Movie Reviews used in Section 4.3.1, the same sentiment classifiers are used. The obtained results are shown in the following table.

		Adjective		*Adjective+*		*Tb 005*		*Tb 01*	
	Total	***P***	***N***	***P***	***N***	***P***	***N***	***P***	***N***
spam	1,000	350	498	502	498	373	623	349	641
ham	3,000	1,161	1,238	1,762	1,238	1,255	1,717	1,159	1,787
Percentages (%)									
spam	1,000	35	50	50	50	37	62	35	64
ham	3,000	39	41	59	41	42	57	39	60

Table 4.16: Sentiment analysis of the Youtube comments

Table 4.16 shows that while in the previous experiments spam messages where more positive than ham messages, in this case, spam comments are more negative than ham comments.

Predictive experiments and comparison. In order to analyze the influence of the sentiment analysis in spam filtering predictive experiments are carried out.

With the objective of identifying the best spam classifiers, several spam classifiers using different settings are applied to the Youtube Comments dataset.

Following the strategy used in Section 4.3.1 7 different classifiers and 56 settings combinations per each classifiers are applied (392 combinations in total), and the ten best results are presented in Table 4.17.

Once the best classifiers are identified, new datasets are created adding the polarity feature to Youtube Comments dataset, using the four sentiment classifiers shown previously.

#	Spam classifier	TP	TN	FP	FN	Acc
1	NBM.c.stwv.go.ngtok	389	2911	89	611	82.50
2	NBMU.c.stwv.go.ngtok	389	2911	89	611	82.50
3	NBM.stwv.go.ngtok	370	2929	71	630	82.48
4	NBMU.stwv.go.ngtok	370	2929	71	630	82.48
5	NBM.c.stwv.go.ngtok.stemmer	379	2919	81	621	82.45
6	NBMU.c.stwv.go.ngtok.stemmer	379	2919	81	621	82.45
7	NBM.stwv.go.ngtok.stemmer	358	2936	64	642	82.35
8	NBMU.stwv.go.ngtok.stemmer	358	2936	64	642	82.35
9	CNB.stwv.go.ngtok	417	2875	125	583	82.30
10	CNB.stwv.go.ngtok.stemmer	400	2891	109	600	82.28

Table 4.17: Results of the best ten classifiers

Then, we apply the best ten classifiers to the labeled datasets and we compare the obtained results with those obtained without polarity feature. The comparison between different results is presented in Tables 4.18 and 4.19. Tables show that sentiment analysis of the texts can help to improve the filtering results using an OSN dataset too. For instance, the best accuracy of the original dataset is improved from an 82.50% to an 82.53% using the polarity feature. Furthermore, the number of false positive are reduced in all the cases, reducing by 10% the original number in some cases (for example, from 89 to 70).

4.4 Conclusions of the Chapter

The main objective of a spam campaign is to sell a product, to trick a user to provide confidential data, or convince a victim to open an attachment. Consequently, is assumed that a special connotation in the message is needed in order to deceive the receivers. We analyzed this assumption using sentiment classifier, and significant differences between spam and legitimate texts were identified.

This chapter shows that it is possible to improve spam filtering adding the polarity of each text to the dataset. We have demonstrated that sentiment analysis of the texts can help to detect spam. In the three different scenarios the results obtained using the polarity feature are better in terms of accuracy, and the number of false positive is reduced.

Further, taking into account that the used sentiment classifiers are independent from the text, it is supposed that the results of a training-based one will be better.

Despite the difference in the percentages of the accuracies does not seem to be relevant, if we take into account the amount of real spam traffic, the improvement is significant.

	Sentiment analyzer								
	None			*Adjective*			*Adjective+*		
#	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	89	611	82.50	70	641	82.23	71	648	82.03
2	89	611	82.50	70	641	82.23	71	648	82.03
3	71	630	82.48	56	657	82.18	55	664	82.03
4	71	630	82.48	56	657	82.18	55	664	82.03
5	81	621	82.45	60	640	82.50	60	643	82.43
6	81	621	82.45	60	640	82.50	60	643	82.43
7	64	642	82.35	54	662	82.10	52	669	81.98
8	64	642	82.35	54	662	82.10	52	669	81.98
9	125	583	82.30	88	615	82.43	79	624	82.43
10	109	600	82.28	75	628	82.43	68	633	82.48

Table 4.18: Comparing original results with the results obtained using own polarity classifiers

	Sentiment analyzer								
	None			*TextBlob005*			*TextBlob01*		
#	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	89	611	82.50	82	625	82.33	83	625	82.30
2	89	611	82.50	82	625	82.33	83	625	82.30
3	71	630	82.48	66	640	82.35	67	640	82.33
4	71	630	82.48	66	640	82.35	67	640	82.33
5	81	621	82.45	74	627	82.48	74	625	82.53
6	81	621	82.45	74	627	82.48	74	625	82.53
7	64	642	82.35	59	652	82.23	59	653	82.20
8	64	642	82.35	59	652	82.23	59	653	82.20
9	125	583	82.30	104	600	82.40	104	600	82.40
10	109	600	82.28	94	612	82.35	94	612	82.35

Table 4.19: Comparing original results with the results obtained using TextBlob polarity classifiers

CHAPTER 5

INFLUENCE OF PERSONALITY RECOGNITION IN SPAM FILTERING

This chapter provides a baseline for a new spam filtering method. The objective is to demonstrate that spam text is written with a specific personality characteristic that can be used to distinguish spam from legitimate text. We hypothesize that being spam a text that generally aims at selling services or products, analyzing its meaning, and specially the personality of the spam text, it can bring similar personality functions such that classification systems are improved.

First, the personality recognition topic is introduced, explaining the meaning and different possibilities offered by personality recognition techniques.

After that, a spam detection method is proposed, and three different types of spam texts are analyzed in order to validate the method: email spam, SMS spam and spam on OSNs. Personality is computed per each spam text, and the features that characterize it are added to the original datasets. Finally, several spam filters with and without personality are compared in terms of accuracy and the number of false positive.

5.1 Relevant Proposals

Personality is a psychological construct aimed at explaining a wide variety of human behaviors in terms of a few, stable and measurable individual characteristics [155]. As authors explain in [22], two main models to formalize personality have been defined so far: the Myers-Briggs personality model [18], which defines the

personality using four dimensions: Extroversion or Introversion, which describes how a person gets energized; Thinking or Felling, which describes the means a person uses to make decisions; Judging or Perceiving, which describes the speed with which a person makes decisions; and Sensing or iNtuition, describes how a person takes in information; and the Big Five model [29] which divides the personality in 5 traits: Openness to experience, Conscientiousness, Extroversion, Agreeableness and Neuroticism.

As it is shown in [101], every text contains information about the personality of the authors, being this the reason that personality recognition became a potential tool for Natural Language Processing. During the last years, different research in personality recognition in blogs [114], offline texts [101] or online social networks [13, 131] have been published.

In [142], authors apply personality recognition to an email feature set. They prove that personality prediction is feasible. This work shows that it is possible to predict the personality of a writer using email messages. While in the literature the personality recognition techniques are applied to more than one text of each user in order to obtain the personality of a certain user, in our case we focus on the personality of each text. We assume that the authors of the texts are unknown, so we can not group the texts of the same author.

Moreover, personality recognition is used in order to detect opinion spam in social media [69], and other researchers present the relationship between personality traits and deceptive communication, which is used to confuse or misleading the user [53].

5.2 Proposed Method

This method has been developed following the procedure of Figure 5.2, which is very similar to the procedure used in Chapter 4. The main difference is that in this case personality recognition techniques are used instead of sentiment analysis techniques.

Taking as a baseline the top classifiers identified in Chapter 4, we analyze the influence of the personality feature in spam filtering. To do that, we compare the results of the best classifiers applied to the dataset with and without personality.

As in the previous chapter, we use a 10-fold cross-validation technique for validation purposes, and the results are analyzed in terms of false positive number and accuracy.

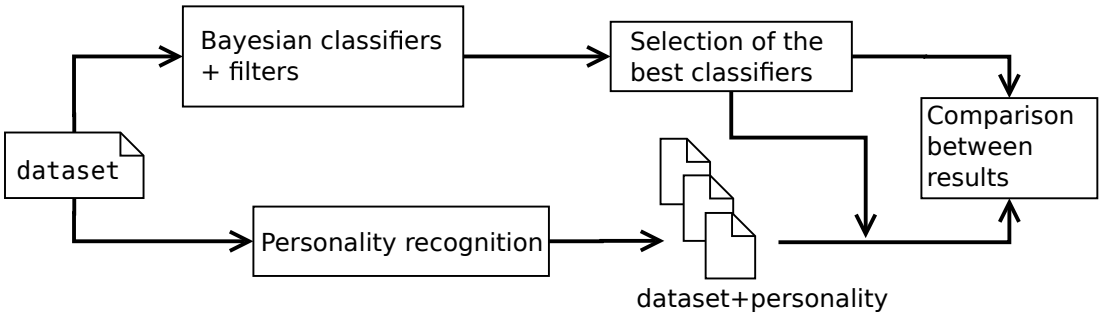

Figure 5.1: Improving spam filtering using personality recognition techniques

5.2.1 Personality recognition

The objective of the next phase is to apply a personality recognition technique to each text in order to add the result as a new feature to the original dataset.

One of the most trusted personality recognition assessment is used in this work: Myers-Briggs personality model. To determine the personality of each text, the four different dimensions of this model are computed: Extroversion/Introversion, Thinking/Feeling, Judging/Perceiving and Sensing/iNtuition. In this case, a publicly available machine learning web service for text classification is used. This service is hosted in *uClassify*[1]. Among all the possibilities offered in this website, we focus on the Myers-Briggs functions developed by Mattias Östmar.

As author explains[2], 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 [78].

Those are the used functions:

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

[1]https://www.uclassify.com
[2]https://www.uclassify.com/browse/prfekt

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 Sensing, the value for iNtuition is 1-X. Thus, we only record one value per each function: Extroversion, Sensing, Thinking and Judging.

In order to create a new dataset containing the personality features, those four values of each email message are added to the original dataset. This new dataset is used during the tests to evaluate the influence of the personality in spam filtering. To do so, we apply the top ten classifiers mentioned previously to both the original dataset and to the new one, and we compare the results.

5.3 Validation of the Proposed Method

In this Section the validation of the proposed method is presented, taking into account the results of three different scenarios.

5.3.1 Email Spam

The first validation of the method is carried out focusing on email spam messages. The method described in Section 5.2 is followed in order to analyze the influence of personality recognition techniques in email spam filtering.

Email datasets used in personality analysis

Two different publicly available datasets are used:

- *CSDMC 2010 Spam Corpus*[3]: composed of 2,949 legitimate email messages and 1,378 spam messages. This dataset is used to carry out the experiments.
- *TREC 2007 Public Corpus*[4]: This corpus contains 75,419 messages: 25,220 ham (legitimate) and 50,199 spam emails. The experiments done with the previous dataset are repeated with this new one in order to validate the results obtained. For the sake of using a similar approach, 4,000 emails are selected randomly from this dataset (3,000 ham and 1,000 spam).

Personality recognition of email messages

In this task, personality recognition functions described in Section 5.2 are used. So the four different dimension of the Myers-Briggs personality model are added to the original dataset.

[3]http://www.csmining.org/index.php/spam-email-datasets-.html
[4]http://plg.uwaterloo.ca/ gvcormac/treccorpus07/

Bayesian spam filtering of email messages

To analyze if personality recognition techniques improve Bayesian spam filtering, the best ten classifiers identified in Section 5.3.1 for spam filtering are used:

1. BLR.i.t.c.stwv.go.wtok
2. DMNB.c.stwv.go.wtok
3. DMNB.i.c.stwv.go.wtok
4. DMNB.i.t.c.stwv.go.wtok
5. DMNB.stwv.go.wtok
6. DMNB.c.stwv.go.stemmer
7. DMNB.i.c.stwv.go.stemmer
8. DMNB.i.t.c.stwv.go.stemmer
9. DMNB.stwv.go.stemmer
10. BLR.i.t.c.stwv.go.ngtok.stemmer.igain

During this chapter, our main objective is to improve those results using the selected classifiers.

Experimental results

The results obtained during the validation of the proposed method are shown. To carry out the following experiments the *CSDMC2010* dataset is used.

Descriptive analysis. The objective of this step is to analyze the personality features of the authors (spammers and legitimate email writers) applying the previously explained personality recognition functions. During this step the personality features are added to the original dataset creating a new labeled dataset, and we extract descriptive statistic regarding the personality. This information is shown in Table 5.1.

Analyzing the data presented in the descriptive table, significant differences between spam and ham can be found. The biggest difference according to Myers-Briggs personality model between spam emails and legitimate emails is given by the Perceiving Function. Taking into account only this dimension, the *sensing* feature of the legitimate emails is 16 points higher than spam emails.

	Total	**Extroversion**	**Sensing**	**Thinking**	**Judging**
ham	2949	975	2439	313	1908
spam	1378	591	918	301	915
Percentage(%)					
ham	100	33	83	11	65
spam	100	43	67	22	66

Table 5.1: Descriptive analysis of the dataset.

This finding can be used to better distinguish between ham and spam, thus classification algorithms may provide better detection rates. In the next steps different experiments are carried out to see the real influence of personality feature in Bayesian email spam filtering.

Using personality. To confirm if personality improves Bayesian spam filtering, we apply the top ten classifiers to the labeled (personality) dataset. We compare the results with those results obtained when applying the same classifiers to the original dataset.

The results obtained during this experiments are shown in Table 5.2. Once again, we use the same nomenclatures as described in Table 1.

		Original dataset			*Personality*		
#	**Spam classifier**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	BLR.i.t.c.stwv.go.wtok	13	24	99.15	14	26	99.08
2	DMNB.c.stwv.go.wtok	21	17	99.12	22	16	99.12
3	DMNB.i.c.stwv.go.wtok	21	17	99.12	22	16	99.12
4	DMNB.i.t.c.stwv.go.wtok	21	17	99.12	22	16	99.12
5	DMNB.stwv.go.wtok	21	17	99.12	22	16	99.12
6	DMNB.c.stwv.go.stemmer	22	19	99.05	22	21	99.01
7	DMNB.i.c.stwv.go.stemmer	22	19	99.05	22	21	99.01
8	DMNB.i.t.c.stwv.go.stemmer	22	19	99.05	22	21	99.01
9	DMNB.stwv.go.stemmer	22	19	99.05	22	21	99.01
10	BLR.i.t.c.stwv.go. .ngtok.stemmer.igain	14	28	99.03	13	26	99.10

Table 5.2: Comparison between normal and personality

Results show that only in one case the classification of the original dataset is improved using personality features (from 99.03% to 99.10%), while in other four cases we obtain the same accuracy (99.12%) and in the other five the accuracy is reduced.

As a conclusion, adding the four personality dimensions to the dataset does not improve the classification in general terms. But as seen in Table 5.1, the Sensing/iNtuition dimension of the personality can be sufficient to better distinguish between the two classification labels.

Myers-Briggs Perceiving Function. To see if the mentioned dimension affects in the Bayesian spam filtering, a new dataset is created. We only use the Myers-Briggs Perceiving Function in order to add the *sensing* characteristic of each message to the dataset. This function has been selected taking into account the differences in personality dimensions, between ham and spam texts, presented in the descriptive analysis of the dataset. Table 5.1 shows that the biggest difference (83% vs 67%) was observed in this dimension.

The followed procedure is the same than in the previous experiment: we apply the best ten classifiers to the new dataset and we compare the results with the original ones.

Table 5.3 summarizes the results obtained.

	Original dataset			*Sensing*		
#	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	13	24	99.15	15	27	99.03
2	21	17	99.12	21	17	99.12
3	21	17	99.12	21	17	99.12
4	21	17	99.12	21	17	99.12
5	21	17	99.12	21	17	99.12
6	22	19	99.05	22	18	99.08
7	22	19	99.05	22	18	99.08
8	22	19	99.05	22	18	99.08
9	22	19	99.05	22	18	99.08
10	14	28	99.03	14	26	99.08

Table 5.3: Results using *sensing* feature

In this case, we obtain better results in terms of accuracy than using all the dimensions of the Myers-Briggs personality model. The results are improved in five cases, in four of them the same results are obtained, and only in one case the result the accuracy is decreased.

Those results give a baseline to see the possibilities that personality recognition techniques can improve Bayesian spam filtering.

Second dataset. Once the mentioned results are obtained, we repeat the process using a different dataset in order to validate the assumptions. As in the previous

Chapter, the TREC2007 dataset is used.

On the one hand, the same ten classifiers that obtained the best results with the previous dataset are applied to TREC2007.

On the other hand, the same classifiers are applied to a new dataset, which has been created adding personality features to the original dataset. Also in this case, the previously used personality recognition functions are used.

Table 5.4 shows the obtained results.

		Original dataset			*Personality*		
#	**Spam classifier**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	BLR.i.t.c.stwv.go. .ngtok.stemmer.igain	17	24	98.98	17	18	**99.13**
2	DMNB.c.stwv.go.stemmer	21	21	98.95	22	19	98.98
3	DMNB.i.c.stwv.go.stemmer	21	21	98.95	22	19	98.98
4	DMNB.i.t.c.stwv.go.stemmer	21	21	98.95	22	19	98.98
5	DMNB.stwv.go.stemmer	21	21	98.95	22	19	98.98
6	DMNB.c.stwv.go.wtok	21	23	98.90	20	23	98.93
7	DMNB.i.c.stwv.go.wtok	21	23	98.90	20	23	98.93
8	DMNB.i.t.c.stwv.go.wtok	21	23	98.90	20	23	98.93
9	DMNB.stwv.go.wtok	21	23	98.90	20	23	98.93
10	BLR.i.t.c.stwv.go.wtok	22	28	98.75	21	27	98.80

Table 5.4: Comparison of the best ten classifiers, second dataset

In this case, although the number of false positive is reduced only in four out of ten classifiers, the accuracy is improved in all of them. Moreover, the best accuracy is improved from 98.98% to 99.13%. Those results provide means to demonstrate that personality recognition helps in email spam filtering.

5.3.2 SMS Spam

This study has been carried out following the procedure showed in Section 5.2, where two main phases are defined: on the one hand, personality recognition techniques are applied to the dataset in order to create a new one when aggregating the personality-based features. This new dataset is later used to see the real influence of the personality feature in SMS spam filtering.

On the other hand, two experiments are carried out applying several classifiers to the datasets (with and without personality) to compare all the results and see which classifies better. Once we analyze the results, the same experiments using a different dataset are carried out. In other words, we use a new labeled (spam/legitimate) SMS dataset, we add the personality to each message, and we apply the same classifiers and the same filters. Finally we analyzed all the results.

SMS datasets used in personality analysis

During this work two publicly available dataset are used:

- *SMS Spam Collection v.1*[5] (named as SMSSpam in this dissertation): Published in [7]. It is composed of 5,574 English, real and non-enconded messages, tagged as being 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*[6] (named as BritishSMS in this dissertation): Introduced in [113]. 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.

Personality recognition of SMS messages

The first phase in our method aims to apply personality recognition techniques to each SMS message in order to create a new dataset, adding this feature to the original dataset.

To do that, the same personality recognition functions described in Section 5.2 are used. So the four different dimension of the Myers-Briggs personality model are added to the original dataset.

SMS spam filtering

To analyze the influence of the personality of SMS messages, we first select 10 representative classifiers and some of the most used filters settings for natural language processing. To do the selection we take into account the results obtained in 5.3.1 and in other research studies such as [91]. Used classifiers: Large-scale Bayesian logistic regression for text categorization, discriminative parameter learning for Bayesian networks, complement class Naive Bayes classifier, Multi-nominal Naive Bayes classifier, updateable multi-nominal Naive Bayes classifier, decision tree (C4.5), random tree, forest of random trees, Support Vector Machine (SMO), and adaptive boosting meta-algorithm with Naive Bayes..

In the same way as in Section 4.3.2, the next step is to apply those classifiers, combined with the best three filters and settings, to the datasets (with and without personality) and compare the results.

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

This first step provides the best classifier for text messages. In the following phase, the best classifiers are selected. We apply those classifiers with different combination of filters and settings (56 combinations per classifier) to the datasets as a second dataset. The objective of the combination of these filters and settings is to follow a text mining process in order to compare results and identify the best ones.

Once the process is evaluated with the SMSSpam dataset, we repeat them using the BritishSMS dataset to compare results, in order to validate our hypothesis.

As in the previous experiments, the nomenclatures of Table 1 are used to represent classification algorithms and filters.

Experimental results

In this Section, results obtained during the previously explained experiments are shown. During the study, we use SMSspam dataset in the first two experiments and BritishSMS to validate the results.

Descriptive analysis. The objective of this step is to analyze the personality of the messages applying the previously described personality recognition technique. The statistics about the personality in SMS messages are computed, and results shown in Table 5.5.

	Total	**Extroversion**	**Sensing**	**Thinking**	**Judging**
ham	4827	4392	3998	2431	1793
spam	747	599	566	238	431
Percentage(%)					
ham	100	91	83	50	37
spam	100	80	76	32	58

Table 5.5: Descriptive analysis of the dataset

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 writen (spam/ham) varies from a personality perspective.

SMS spam filtering classifiers that perform better. These experiments aim to identify the best SMS spam filtering classifiers in order to use them in the next experiment with more filters and settings. As it is mentioned previously, we choose 10 classifiers and three different filter combinations per each classification technique.

The same classifiers and settings used in Section 4.3.2 are selected in order to follow the same procedure.

Table 5.6 shows the results of the ten classifiers in combination with the three filters. Numbers represent the type of filters used, namely: (1) *stwv.go.wtok*, (2) *i.t.c.stwv.go.ngtok.stemmer.igain* and (3) *i.t.c.stwv.go.wtok*.

SMO technique appears in the Table as the classifier that better results provide, when used with the third filter settings. Although in this case the personality does not provide means to improve the original classification scores, the other two cases that used personality features reach the same accuracy. Additionally, it can be seen that in most of the cases, accuracy is improved, or at least same results are obtained.

Table 5.6 also shows that the number false positive is reduced generally. This means that personality can help to improve the results in SMS spam filtering, but we need to analyze more possible combinations of filter settings to confirm this statement.

SMS spam filtering with personality features. The second experiment is based on the results obtained in the first one. While the previous aims to search for the best algorithms, this one aims to explore the possible filter combinations with the best classifiers.

In this manner, we identify the best 6 classifiers from Table 5.6 and combine each one with 56 different filter settings. We analyze the achieved results and we select the classifiers that obtained the best ten results in terms of accuracy. Finally we apply those classifiers to the new dataset that we created using the personality recognition technique. Those results are shown in Table 5.7.

The table shows that a higher accuracy than in the previous experiment is obtained applying new settings of the filters to the original SMS dataset.

Analyzing the information shown in Table 5.7 we see that the aggregation of the personality feature improves almost all of the results. In terms of accuracy a 98.94% is reached improving the best result obtained of the original dataset. Only in two cases the accuracy does not improve, but the number of false positive is reduced in both (from 27 to 19 and from 13 to 3). In addition, the number of false positive is reduced in all cases. For the spam problem, reduction of the false positives can be considered a significant improvement.

	Original dataset		*Personality*	
Spam classifier	**FP**	**Acc**	**FP**	**Acc**
SMO.3	3	98.73	5	98.64
NBM.3	12	98.69	4	98.71
NBMU.3	12	98.69	3	98.42
BLR.3	5	98.64	5	98.64
DMNB.1	10	98.62	5	98.73
BLR.2	2	98.60	2	98.60
NBM.1	23	98.53	14	98.62
SMO.2	4	98.53	3	98.56
NBMU.2	36	98.51	18	98.62
NBMU.1	29	98.49	19	98.64
CNB.1	31	98.44	18	98.60
NBM.2	52	98.37	44	98.51
DMNB.2	4	98.28	2	98.31
DMNB.3	4	98.28	2	98.31
CNB.2	64	98.19	54	98.35
CNB.3	56	98.17	23	98.73
BLR.1	1	97.45	0	96.23
SMO.1	0	97.45	0	97.43
J48.3	54	97.02	57	96.97
J48.2	58	96.90	60	96.86
J48.1	42	96.86	43	96.91
RF.2	0	96.38	0	96.47
RF.3	0	96.21	0	96.23
RT.1	25	95.60	22	95.39
RF.1	0	95.19	0	94.83
RT.3	84	95.16	86	94.94
RT.2	88	95.07	98	94.73
AB.2	167	91.44	167	91.44
AB.3	167	91.44	167	91.44
AB.1	188	91.32	188	91.32

Table 5.6: Comparison between results of the first experiment of SMS spam filtering

		Original dataset			*Personality*		
#	**Spam classifier**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	NBMU.i.c.stwv.go.ngtok	28	36	98.85	26	39	98.83
2	NBMU.i.t.c.stwv.go.ngtok	27	39	98.82	19	40	**98.94**
3	NBM.i.t.c.stwv.go.ngtok	32	36	98.78	28	36	98.85
4	NBMU.i.t.c.stwv.go.ngtok.stemmer	23	45	98.78	19	48	98.80
5	NBM.c.stwv.go.wtok	13	56	98.76	5	63	98.78
6	NBM.i.t.c.stwv.go.ngtok.stemmer	34	35	98.76	30	38	98.78
7	NBMU.c.stwv.go.wtok	13	56	98.76	**3**	81	98.49
8	CNB.i.t.c.stwv.go.ngtok.stemmer	37	34	98.73	34	35	98.76
9	NBM.i.c.stwv.go.ngtok	37	34	98.73	33	35	98.78
10	NBM.i.c.stwv.go.ngtok.stemmer	36	35	98.73	33	36	98.76

Table 5.7: Comparison of the best ten classifiers

Validation of the results using a second dataset. As we mentioned previously, so as to validate the results during the first two experiments, we decided to repeat those experiments with a different dataset. In this case, we use the publicly available BritishSMS dataset.

To accomplish the validation of the experiment, we select the same ten classifiers that obtained the best results with the previous 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 5.8.

		Original dataset			*Personality*		
#	**Spam classifier**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	NBM.i.c.stwv.go.ngtok	5	17	97.49	5	17	97.49
2	NBM.i.t.c.stwv.go.ngtok	5	18	97.37	**3**	19	**97.49**
3	NBMU.i.c.stwv.go.ngtok	7	16	97.37	6	16	97.49
4	NBMU.i.t.c.stwv.go.ngtok	6	17	97.37	6	17	97.37
5	NBMU.i.t.c.stwv.go.ngtok.stemmer	9	16	97.14	9	17	97.03
6	NBM.i.c.stwv.go.ngtok.stemmer	8	18	97.03	7	18	97.14
7	CNB.i.t.c.stwv.go.ngtok.stemmer	8	19	96.91	6	19	97.14
8	NBM.i.t.c.stwv.go.ngtok.stemmer	8	19	96.91	6	19	97.14
9	NBM.c.stwv.go.wtok	9	23	96.34	2	26	96.80
10	NBMU.c.stwv.go.wtok	9	23	96.34	9	22	96.46

Table 5.8: Comparison of the best ten classifiers, second dataset

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

Taking into account that it is a very small SMS spam dataset (875 SMSs), the generalization of the best result would significantly detect a very high amount of spam SMSs.

5.3.3 Social Media Spam

With the objective of demonstrating that spam filtering on OSNs can be improved using personality recognition techniques, the next experiment is done.

To do that, the same procedure explained in Section 5.3.1 is followed but using a different dataset.

Personality recognition techniques are used to add the personality feature to the dataset, in the same time that a descriptive analysis of the texts is done. After that, several classifiers and filters settings are combined and applied to the dataset with and without personality. And finally, the obtained results are compared in order to give conclusions about the study.

OSN spam dataset used in personality analysis

The dataset used in this experiment is the same that we present in the previous chapter:

Youtube Comments Dataset [7]: Presented in [115]. This dataset contains multilingual 6,431,471 comments from a popular social media website Youtube, and 481,334 are marked as spam. Among all those comments the subset created in Section 4.3.3 is used here.

Experimental results

This phase is divided in two main parts: on the one hand, the descriptive experiment of the dataset, and on the other hand, the predictive experiments and the comparison between the results.

Descriptive experiment. Taking into account the personality recognition functions presented in the previous sections, a descriptive analysis of the dataset is done. During this experiment, the different dimensions of the personality model are added to the original dataset, and a new dataset is created.

Although the differences between ham and spam comments are not significant, Table 5.9 shows that the biggest difference is in terms of *thinking* feature. So in the

[7] http://mlg.ucd.ie/yt/

	Total	**Extroversion**	**Sensing**	**Thinking**	**Judging**
ham	3000	1841	2222	2303	1651
spam	1000	624	715	726	538
Percentage(%)					
ham	100	61	74	77	55
spam	100	62	72	73	54

Table 5.9: Descriptive analysis of the dataset

next step, first of all a experiment using all the dimensions is carried out and after that, also a test is done adding only the *thinking* feature to the original dataset.

Predictive experiments and comparison. To analyze the if personality recognition techniques help in OSNs spam filtering, the same combinations presented in Section 5.3.1 are used in order to identify the best ten classifiers. The same test was carried out in the previous chapter, so we know which are the best ten classifiers.

Knowing that, the next step is to apply those classifier to the labeled dataset in order to compare the results.

		Original dataset			*Personality*		
#	**Spam classifier**	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	NBM.c.stwv.go.ngtok	89	611	82.50	51	663	82.15
2	NBMU.c.stwv.go.ngtok	89	611	82.50	43	678	81.98
3	NBM.stwv.go.ngtok	71	630	82.48	42	679	81.98
4	NBMU.stwv.go.ngtok	71	630	82.48	32	699	81.73
5	NBM.c.stwv.go.ngtok.stemmer	81	621	82.45	46	665	82.23
6	NBMU.c.stwv.go.ngtok.stemmer	81	621	82.45	37	683	82.00
7	NBM.stwv.go.ngtok.stemmer	64	642	82.35	39	688	81.83
8	NBMU.stwv.go.ngtok.stemmer	64	642	82.35	29	707	81.60
9	CNB.stwv.go.ngtok	125	583	82.30	60	646	82.35
10	CNB.stwv.go.ngtok.stemmer	109	600	82.28	54	650	82.40

Table 5.10: Comparison of the best ten classifiers

Results in Table 5.10 show that while the number of false positive is reduced in every cases, the accuracy is only improved in two out of ten cases.

In this point, taking into account the results obtained in the descriptive experiment, where we can see that the main difference between ham and spam comments is the *thinking* feature, the experiment is repeated but adding only this dimension

to the original dataset. The results obtained during this experiment are presented in Table 5.11.

	Original dataset			*Thinking*		
#	**FP**	**FN**	**Acc**	**FP**	**FN**	**Acc**
1	89	611	82.50	76	629	82.38
2	89	611	82.50	70	633	82.43
3	71	630	82.48	61	645	82.35
4	71	630	82.48	56	650	82.35
5	81	621	82.45	69	632	82.48
6	81	621	82.45	65	636	82.48
7	64	642	82.35	56	648	82.40
8	64	642	82.35	52	657	82.28
9	125	583	82.30	100	608	82.30
10	109	600	82.28	87	615	82.45

Table 5.11: Comparison of the best ten classifiers with and without *Thinking* dimension

In this case, the accuracy is improved in more classifiers than in the previous table. The number of false positives is also reduced compared to the original dataset. Moreover, the best accuracy (82.50%) is not improved but the same percentage is obtained.

The significant reduction of the number of false positive give means to validate the hypothesis that personality recognition techniques help in OSNs spam filtering.

5.4 Conclusions of the Chapter

This Chapter provides means to give a baseline for improving spam filtering.

This study shows that adding a personality score obtained during a personality analysis of texts in most of the cases the result is improved in both terms: accuracy and the number of false positive.

Moreover, in order to validate the results and to demonstrate that the personality recognition is useful in spam filtering, we carry out the same experiment using different datasets. The results in both experiments are positive, improving the accuracy and the number of false positives of most of the best classifiers.

Furthermore, taking into account that the personality recognition functions used are independent from the text, the use of manually tagged (personality) emails during the learning process of the function might improve the results.

Although the difference in percentage does not seem to be relevant, taking into account the amount of real spam traffic, the improvement is significant.

CHAPTER 6

COMBINATION OF SENTIMENT ANALYSIS AND PERSONALITY RECOGNITION IN SPAM FILTERING

The main objective of this chapter is to explain a new spam filtering method, demonstrating the combination of sentiment analysis and personality techniques applied to spam, can improve spam classification. We take into account the results obtained in Chapters 4 and 5, aiming to impruve previous results.

The details of the proposed method are described first, and validated later using the same datasets (CSDMC2010, TREC2007, SMSSpam, BritishSMS, and Youtube) of the previous two Chapters.

6.1 Proposed Method

The proposed method is a combination of the two previously described processes. As shown in Figure 6.1, each original dataset is fed with sentiment, and personality features in a way that four datasets are kept for comparison: the original one, the original with a polarity feature, the original with the personality feature, and finally the aggregation of both polarity and personality features to the original dataset. Second, the 10 classifiers that better discriminate the datasets (CSDMC2010, TREC2007, SMSSpam, BritishSMS, and Youtube) are identified. Finally those classifiers are applied to all the datasets in order to compare the results.

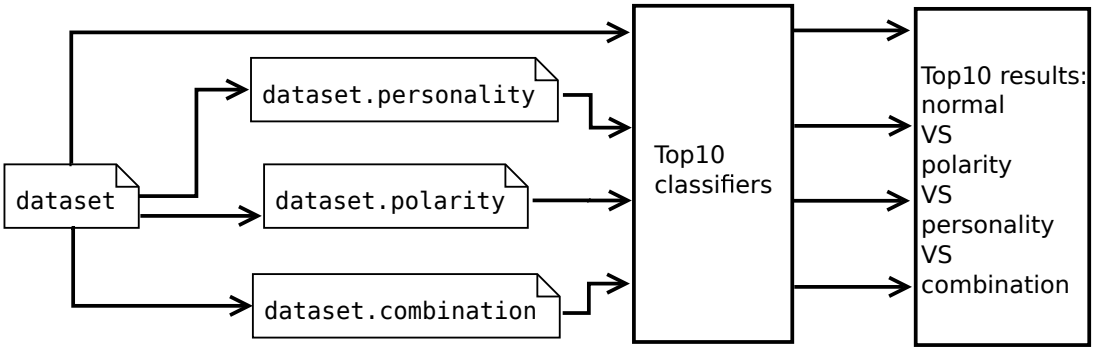

Figure 6.1: Improving spam filtering combining sentiment analysis and personality recognition

6.2 Validation of the Proposed Method

Once again, three different scenarios are analyzed in order to conduct the comparison: email spam, SMS spam and social media spam.

6.2.1 Email Spam

As in the previous chapters, we use CSDM2010 and TREC2007 datasets. In both of them the best sentiment analyzer and the best personality dimensions are selected, in order to compare the best results of Chapters 4 and 5 with the new results obtained using the combined dataset. The best sentiment analyzer and the best personality technique are also used to create the new dataset with both features.

In Table 6.1, the results of the first experiments are shown. In the first column the results of using the original dataset are presented, while in the second one the best results obtained using sentiment analysis are summarized. In the third one, the best results of this dataset using personality detection techniques are given. Finally, the results obtained with the combination of personality and polarity are given.

According to the obtained results, we can confirm the combination of sentiment analysis with personality recognition techniques improved the best result obtained in Bayesian spam filtering in terms of accuracy. The combination improves (with a 99.24% of accuracy) both the top result of the original dataset (99.15%) and the top result of the polarity analysis (99.21%). Moreover, in those cases where the best result is achieved, the combination of sentiment analysis and personality techniques reduces the number false positive.

To validate those first results, the same test is carried out using the TREC2007 dataset and the obtained results are summarized in Table 6.2.

	Used technique							
	None		*TB 0.1*		*Sensing*		*Comb*	
Spam classifier	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**
BLR.i.t.c.stwv.go.wtok	13	**99.15**	14	99.12	15	99.03	15	99.03
DMNB.c.stwv.go.wtok	21	99.12	22	99.21	21	99.12	19	**99.24**
DMNB.i.c.stwv.go.wtok	21	99.12	22	99.21	21	99.12	19	**99.24**
DMNB.i.t.c.stwv.go.wtok	21	99.12	22	99.21	21	99.12	19	**99.24**
DMNB.stwv.go.wtok	21	99.12	22	99.21	21	99.12	19	**99.24**
DMNB.c.stwv. .go.stemmer	22	99.05	22	**99.15**	22	99.08	23	99.05
DMNB.i.c.stwv. .go.stemmer	22	99.05	22	**99.15**	22	99.08	23	99.05
DMNB.i.t.c.stwv. .go.stemmer	22	99.05	22	**99.15**	22	99.08	23	99.05
DMNB.stwv.go.stemmer	22	99.05	22	**99.15**	22	99.08	23	99.05
BLR.i.t.c.stwv.go. .ngtok.stemmer.igain	14	99.03	14	99.03	14	99.08	14	**99.10**

Table 6.1: Comparison of the best classifiers using the dataset CSDMC2010

	Used technique							
	None		*Adj*		*Pers*		*Comb*	
Spam classifier	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**
BLR.i.t.c.stwv.go. .ngtok.stemmer.igain	17	98.98	17	99.05	17	99.13	17	**99.18**
DMNB.c.stwv. .go.stemmer	21	98.95	20	99.05	22	98.98	21	**99.10**
DMNB.i.c.stwv. .go.stemmer	21	98.95	20	99.05	22	98.98	21	**99.10**
DMNB.i.t.c.stwv. .go.stemmer	21	98.95	20	99.05	22	98.98	21	**99.10**
DMNB.stwv.go.stemmer	21	98.95	20	99.05	22	98.98	21	**99.10**
DMNB.c.stwv.go.wtok	21	98.90	20	98.93	20	98.93	21	**98.95**
DMNB.i.c.stwv.go.wtok	21	98.90	20	98.93	20	98.93	21	**98.95**
DMNB.i.t.c.stwv.go.wtok	21	98.90	20	98.93	20	98.93	21	**98.95**
DMNB.stwv.go.wtok	21	98.90	20	98.93	20	98.93	21	**98.95**
BLR.i.t.c.stwv.go.wtok	22	98.75	22	98.68	21	98.80	22	**98.85**

Table 6.2: Comparison of the best classifiers using the dataset TREC2007

In this case, the best result of the original dataset is improved using only sentiment analysis, and a better accuracy if obtained using personality detection techniques. Moreover, the combined dataset improves even more all the previous accuracies of each classifier, reaching 99.18% of accuracy.

6.2.2 SMS Spam

The second scenario aims to analyze the method proposed in this Chapter using datasets composed by SMS messages. The same dataset used in the previous Chapters are used also in this case: SMSSpam and BritishSMS.

Table 6.3 shows the results obtained with the first dataset. The sentiment analyzer *TextBlob -0.05* and all the dimensions of the personality recognition model are used to create the combined dataset, and the Tables 6.3 summarizes all the results.

	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

Table 6.3: Comparison of the best classifiers using the dataset SMSSpam

Best accuracy results of the original dataset (98.85%), are improved with sentiment analysis (98.91%), personality features (98.94%), reaching to a 99.01% with the combination of the features.

In addition, to validate those results a second dataset is used and the results are shown in Table 6.4. Once again, the best result is improved, obtaining a 97.6% of accuracy, and reducing the number of false positive in most of the classifiers.

	Used technique							
	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

Table 6.4: Comparison of the best classifiers using the dataset BritishSMS

	Used technique							
	None		*TB 0.1*		*Thinking*		*Comb*	
Spam classifier	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**	**FP**	**Acc**
NBM.c.stwv.go.ngtok	89	**82.50**	83	82.30	76	82.38	71	82.30
NBMU.c.stwv.go.ngtok	89	**82.50**	83	82.30	70	82.43	66	82.30
NBM.stwv.go.ngtok	71	**82.48**	67	82.33	61	82.35	57	82.20
NBMU.stwv.go.ngtok	71	**82.48**	67	82.33	56	82.35	51	82.23
NBM.c.stwv.go. .ngtok.stemmer	81	82.45	74	**82.53**	69	82.48	60	82.48
NBMU.c.stwv.go. .ngtok.stemmer	81	82.45	74	82.53	65	82.48	53	**82.55**
NBM.stwv.go. .ngtok.stemmer	64	82.35	59	82.20	56	**82.40**	51	82.18
NBMU.stwv.go. .ngtok.stemmer	64	**82.35**	59	82.20	52	82.28	46	82.13
CNB.stwv.go.ngtok	125	82.30	104	82.40	100	82.30	84	**82.50**
CNB.stwv.go. .ngtok.stemmer	109	82.28	94	82.35	87	**82.45**	75	82.43

Table 6.5: Comparison of the best classifiers using the dataset of Youtube comments

6.2.3 Social Media Spam

Finally, to demonstrate if this new detection method could also be valid in OSN spam, a new experiment is performed using the Youtube comments dataset. The same methodology used in the previous experiments is followed and the different results are presented in Table 6.5.

Yet again, the combination of different techniques improves the results in both terms: accuracy and the number of false positive. Here, the number of false positive is reduced in every case, and the best accuracy is obtained using the combined dataset (82.55%).

6.3 Conclusions of the Chapter

This chapter presents a new filtering method that gives the research community the opportunity of detecting non evident intent in spam. This new method consists in using a combination of the polarity and personality features of each text.

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

Moreover, this method is validated in three different types of spam, and using more than one dataset in each type. This means that although each spam type is different, sentiment analysis and personality recognition techniques are capable to highlight differences between spam and ham texts. Those differences help classifiers to filter spam texts, and to improve the results.

Furthermore, this experiment demonstrates that the more information about the content of the texts is added to the dataset, the improvement of the results is higher.

CHAPTER 7

CONCLUSION

7.1 Summary

In this dissertation we have analyzed the threat of spam from different points of view.

We aimed to alert users of the risk of publicizing personal data in OSNs. The main reason is that all the information that is (intentionally or not) publicly available in OSNs can be used to carry out malicious activities. In this thesis, we have demonstrated the possibility of improving current spam click-through rates using ONS public information to personalize spam messages. We extracted public information from Facebook profiles, we designed email templates analyzing the collected information, and we carried out several experiments sending both typical and personalized spam. The tests provide means to validate our hypothesis about the effectiveness of using OSN information in order to obtain higher click-through rates in spam campaigns.

In the second part, we presented three novel spam detection methods with the objective of improving spam filtering techniques.

Taking into account that the main objective of a spam campaign is to sell a product, to trick a user to provide confidential data, or convince a victim to open an attachment, we assume that those messages need have a special connotation regarding. We analyzed this assumption using different spam datasets (with different types of spam) and applying different sentiment classifiers. We found significant differences between legitimate and spam messages, so we designed a novel filtering method. This method improves spam filtering and reduces the number of false

positives. We validated those results using three different scenarios: email spam, SMS spam and OSN spam. The detection capability is improved in all cases.

Following a very similar procedure, we presented the second spam filtering method. In this case, personality recognition techniques are used instead of sentiment analysis. We consider this technique as a further step towards identifying intentionality of the messages. The personality dimensions of a well known personality model are used during the study. We analyzed the differences between spam and legitimate text, and we took them into account to add the distinguishing features to the original dataset. Also, in this case the three scenarios are considered in the validation and the hypothesis are proved with better detection ratios.

Finally, we combined the two previously mentioned strategies and we presented another method, a combined one. We added both features to the datasets, and we carried out the validation experiments with and without the features. With this combination we provided means to validate our hypothesis that those kind of techniques can help to detect non-evident spam texts. In consequence of using this method, it is possible to identify some insights of the intention of the texts, and more spam texts are correctly classified.

7.2 Future Work

Assuming that our study has some limitations, in this section we outline research directions that can be explored in future work. During this thesis work we observe that our methods offer more possibilities than the ones described in this dissertation, so we explain the most relevant future work in the following lines. Even if we discarded them for the completion of our research, they still deserve the attention of the scientific community.

Personalized spam. Open questions remain in this direction, like the influence of the spam templates in the users, massive targeted spam delivery, or investigating spam campaigns inside the OSN framework. In this dissertation an analysis of the effectiveness of personalized spam is carried out. To do that, we create a email spam campaign and we analyze the results. But, considering the huge increase of spam in OSNs, it would be interesting to analyze personalized spam campaigns and their response rates inside a OSN. This research alternative offers a different vision about users behavior regarding targeted attacks, specially because the information from the same OSN can be used to fully personalize messages. The influence of the spam templates in the users could also be further analyzed, because we understand that the style, content, and writing style can provide different results. Massive targeted spam delivery remains as another research direction; due to legal and

ethical issues we did not go further in spam message campaigns, and used a set of 2,889 emails. This is far from the millions that spammers send.

Sentiment analyzers. In order to improve the accuracy of the sentiment analyzers, we propose to explore the possibility of developing and using a learning-based classifier instead of a lexicon-based classifiers (those based on lexicon resources such as SentiWordNet). This gives the opportunity to analyze the influence of the improvement of the classifiers accuracy on spam filtering results.

Specific calibration. Another attractive direction to enrich the methods presented in this thesis is to use specific calibration datasets in the sentiment analyzer and also in personality recognition functions. For instance, the personality recognition functions used are trained with independent text. The use of emails, SMSs or comments tagged in terms of personality during the learning process of the functions might improve them. The same happens in the case of the proposed learning-based sentiment classifiers, which will offer better accuracies calibrating with specific text pieces.

Content analysis techniques. The use of sentiment analysis and personality recognition techniques give a baseline for future research studies focused on spam filtering using content analysis techniques. Those methods could benefit from deeper linguistic features for classification, which could be combined with sentiment and personality. An example of such features is used in [104], where authors present a multilingual native language identification. The target of our research was identifying the intentionality of a message. We reached to a point where polarity and personality are extracted, but further research should be done in finding intentionality. A corpora of intentionality related words could be generated, and used to compute it, in a similar way of SentiWordNet. SentiWordNet is the result of the automatic annotation of all the synsets of WordNet according to the notions of positivity, negativity, and neutrality. A similar approach with intentionality notions could be pursued.

Identification of spam types. The used datasets contain several types of spam such as advertisements, scams, Nigerian scams... We think that it would be possible to improve the spam filtering results identifying the different spam types before the filtering, due to a specific writing style of each spam type. For instance, a communication with the purpose of selling a product will be positive, while the objective of Nigerian scams is to make the reader feel pity.

BIBLIOGRAPHY

[1] Extracto entregable e2004: "desarrollos de optenet relacionados con el filtrado de spam.". Technical report, OPTENET, S.A., 2011.

[2] Sms and mobile messaging attacks. Technical report, GSMA Spam Reporting Service, January 2011.

[3] Social network security: A syssec whitepaper. Technical report, SysSec Consortium, June 2012.

[4] Sms: the language of 6 billion people. Technical report, Portio Research Limited, June 2015.

[5] Pyzor. `http://sourceforge.net/apps/trac/pyzor/`, 2016.

[6] Inc. Alexa Internet. Alexa top 500 global sites. `http://www.alexa.com/topsites`, 2016.

[7] Tiago A Almeida, José María Gómez Hidalgo, and Akebo Yamakami. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM symposium on Document engineering*, pages 259–262. ACM, 2011.

[8] Ion Androutsopoulos, John Koutsias, Konstantinos Chandrinos, Georgios Paliouras, and Constantine D. Spyropoulos. An evaluation of naive bayesian anti-spam filtering. *CoRR*, cs.CL/0006013, 2000.

[9] Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos, and Constantine D. Spyropoulos. An experimental comparison of naive bayesian and

keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 160–167, New York, NY, USA, 2000. ACM.

[10] Jeremy M Anglin, George A Miller, and Pamela C Wakefield. Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, pages i–186, 1993.

[11] Shlomo Argamon, Moshe Koppel, and Galit Avneri. Routing documents according to style. In *In Proceedings of First International Workshop on Innovative Information Systems*, 1998.

[12] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[13] Shuotian Bai, Tingshao Zhu, and Li Cheng. Big-five personality prediction based on user behaviors at social network sites. *CoRR*, abs/1204.4809, 2012.

[14] Marco Balduzzi, Christian Platzer, Thorsten Holz, Engin Kirda, Davide Balzarotti, and Christopher Kruegel. Abusing social networks for automated user profiling. In *Proceedings of the 13th international conference on Recent advances in intrusion detection*, RAID'10, pages 422–441, Berlin, Heidelberg, 2010. Springer-Verlag.

[15] BBC. Spam reaches 30-year anniversary. `http://news.bbc.co.uk/2/hi/7380788.stm`, May 2008.

[16] Joseph Bonneau, Jonathan Anderson, and George Danezis. Prying data out of a social network. *International Conference on Advances in Social Network Analysis and Mining*, 0:249–254, 2009.

[17] Danah M. Boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 2007. http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html.

[18] Isabel Briggs Myers and Peter B Myers. Gifts differing: Understanding personality type, 1980.

[19] Jaime Guillermo Carbonell. Subjective understanding: Computer models of belief systems. Technical report, DTIC Document, 1979.

[20] Claire Cardie, Janyce Wiebe, Theresa Wilson, and Diane Litman. Combining low-level and summary representations of opinions for multi-perspective question answering. In *In Working Notes - New Directions in Question Answering (AAAI Spring Symposium Series)*, pages 20–27, 2003.

[21] Xavier Carreras and Lluís Màrquez. Boosting trees for anti-spam email filtering. *CoRR*, cs.CL/0109015, 2001.

[22] Fabio Celli and Massimo Poesio. PR2: A language independent unsupervised tool for personality recognition from text. *CoRR*, abs/1402.2796, 2014.

[23] Gobinda G. Chowdhury. Natural language processing. *ARIST*, 37(1):51–89, 2003.

[24] Zi Chu, Indra Widjaja, and Haining Wang. Detecting social spam campaigns on twitter. In Feng Bao, Pierangela Samarati, and Jianying Zhou, editors, *ACNS*, volume 7341 of *Lecture Notes in Computer Science*, pages 455–472. Springer, 2012.

[25] William W. Cohen. Fast effective rule induction. In *In Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995.

[26] William W. Cohen and Haym Hirsh. Joins that generalize: Text classification using whirl. In *In Proc. of the Fourth Int'l Conference on Knowledge Discovery and Data Mining*, pages 169–173, 1998.

[27] Gordon V Cormack. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455, 2007.

[28] Gordon V. Cormack, José María Gómez Hidalgo, and Enrique Puertas Sánz. Spam filtering for short messages. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 313–320, New York, NY, USA, 2007. ACM.

[29] Paul T Costa and Robert R McCrae. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5, 1992.

[30] Sanjiv Das and Mike Chen. Yahoo! for amazon: Extracting market sentiment from stock message boards. In *In Asia Pacific Finance Association Annual Conf. (APFA)*, 2001.

[31] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA, 2003. ACM.

[32] Sarah Jane Delany, Mark Buckley, and Derek Greene. Sms spam filtering: methods and data. *Expert Systems with Applications*, 39(10):9899–9908, 2012.

[33] Sarah Jane Delany, Pádraig Cunningham, and Barry Smyth. Ecue: A spam filter that uses machine learning to track concept drift. In *Proceedings of the 2006 Conference on ECAI 2006: 17th European Conference on Artificial Intelligence August 29 – September 1, 2006, Riva Del Garda, Italy*, pages 627–631, Amsterdam, The Netherlands, The Netherlands, 2006. IOS Press.

[34] Maya Dimitrova, Aidan Finn, Nicholas Kushmerick, and Barry Smyth. Web genre visualization. In *In Proc. conference on human factors in*, 2002.

[35] Xiaowen Ding, Bing Liu, and Lei Zhang. Entity discovery and assignment for opinion mining applications. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1125–1134, New York, NY, USA, 2009. ACM.

[36] Luca Dini and Giampaolo Mazzini. Opinion classification through information extraction. In *Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining)*, pages 299–310, 2002.

[37] H. Drucker, Donghui Wu, and V. N. Vapnik. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048–1054, Sep 1999.

[38] Jeremy J Eberhardt. Bayesian spam detection. *Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal*, 2(1):2, 2015.

[39] Pedro Fabricio Echeverria Briones, Zoila Veronica Altamirano Valarezo, Alvaro Badir Pinto Astudillo, and Johanna Del Carmen Sanchez Guerrero. Text mining aplicado a la clasificación y distribución automática de correo electrónico y detección de correo spam. 2009.

[40] Manuel Egele, Gianluca Stringhini, Christopher KrÃ$\frac{1}{4}$gel, and Giovanni Vigna. Compa: Detecting compromised accounts on social networks. *Symposium on Network and Distributed System Security (NDSS)*, 2013.

[41] F. Erlandsson, R. Nia, M. Boldt, H. Johnson, and S. F. Wu. Crawling online social networks. In *Network Intelligence Conference (ENIC), 2015 Second European*, pages 9–16, Sept 2015.

[42] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[43] Enaitz Ezpeleta, Urko Zurutuza, and José María Gómez Hidalgo. 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.

[44] Enaitz Ezpeleta, Urko Zurutuza, and José María Gómez Hidalgo. Short messages spam filtering using personality recognition. In *Proceedings of the 4th Spanish Conference in Information Retrieval*, 2016.

[45] Enaitz Ezpeleta, Urko Zurutuza, and José María Gómez Hidalgo. A study of the personalization of spam content using facebook public information. *Logic Journal of IGPL*, 2016. In Press.

[46] Enaitz Ezpeleta, Urko Zurutuza, and José María Gómez Hidalgo. Using personality recognition techniques to improve bayesian spam filtering. *Journal Procesamiento del Lenguaje NaturalNatural*, (57), 2016. In Press.

[47] Enaitz Ezpeleta, Urko Zurutuza, and José María Gómez Hidalgo. An analysis of the effectiveness of personalized spam using online social network public information. In *International Joint Conference - CISIS'15 and ICEUTE'15, 8th International Conference on Computational Intelligence in Security for Information Systems / 6th International Conference on EUropean Transnational Education, Burgos, Spain, 15-17 June, 2015*, pages 497–506, 2015.

[48] Enaitz Ezpeleta, Urko Zurutuza, and José María Gómez Hidalgo. Los spammers no piensan: usando reconocimiento de personalidad para el filtrado de spam en mensajes cortos. In *RECSI*, 2016. In Press.

[49] Enaitz Ezpeleta, Urko Zurutuza, and José María Gómez Hidalgo. Short messages spam filtering using sentiment analysis. In *Text, Speech and Dialogue TSD*, 2016. In Press.

[50] Tom Fawcett. "in vivo" spam filtering: A challenge problem for kdd. *SIGKDD Explor. Newsl.*, 5(2):140–148, December 2003.

[51] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[52] Ronen Feldman and Ido Dagan. Knowledge discovery in textual databases (kdt). In *In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pages 112–117. AAAI Press, 1995.

[53] T. Fornaciari, F. Celli, and M. Poesio. The effect of personality type on deceptive communication style. In *Intelligence and Security Informatics Conference (EISIC), 2013 European*, pages 1–6, Aug 2013.

[54] Norbert Fuhr, Stephan Hartmann, Gerhard Lustig, Michael Schwantner, Konstadinos Tzeras, and Gerhard Knorz. Air/x - a rule-based multistage indexing system for large subject fields. In *PROCEEDINGS OF RIAO'91*, pages 606–623, 1991.

[55] Patxi Galán-García, Carlos Laorden Goméz, and Pablo Garcia Bringas. Towards a more efficient and personalised advertisement content in on-line social networks. In Aparna S. Varde and Fabian M. Suchanek, editors, *PIKM*, pages 95–98. ACM, 2012.

[56] Murthy Ganapathibhotla and Bing Liu. Mining opinions in comparative sentences. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 241–248, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[57] Hongyu Gao, Yan Chen, Kathy Lee, Diana Palsetia, and Alok N. Choudhary. Towards online spam filtering in social networks. In *NDSS*. The Internet Society, 2012.

[58] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y. Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 17th ACM conference on Computer and communications security*, CCS '10, pages 681–683, New York, NY, USA, 2010. ACM.

[59] Rohit Giyanani and Mukti Desai. Spam detection using natural language processing. *International Journal of Computer Science Research & Technilogy*, 1:55–58, August 2013.

[60] Cliff Goddard. *Semantic analysis: A practical introduction*. Oxford University Press, 2011.

[61] Paul Graham. *Hackers and painters - big ideas from the computer age*. O'Reilly, 2004.

[62] John Graham-Cumming. The spammers' compendium. In *Proceedings of the MIT Spam Conference*, 2003.

[63] Srishti Gupta, Payas Gupta, Mustaque Ahamad, and Ponnurangam Kumaraguru. Abusing phone numbers and cross-application features for crafting targeted attacks. *CoRR*, abs/1512.07330, 2015.

[64] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 174–181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

[65] Marti Hearst. Direction-based text interpretation as an information access refinement. *Text-Based Intelligent Systems*, pages 257–274, 1992.

[66] Marti Hearst. What is text mining? *Retrieved February*, 7:2011, 2003.

[67] Marti A Hearst. Text data mining: Issues, techniques, and the relationship to information access. In *Presentation notes for UW/MS workshop on data mining*, 1997.

[68] Marti A. Hearst. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 3–10, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[69] Donato Hernández Fusilier, Manuel Montes-y Gómez, Paolo Rosso, and Rafael Guzmán Cabrera. Detecting positive and negative deceptive opinions using pu-learning. *Inf. Process. Manage.*, 51(4):433–443, July 2015.

[70] Jordi Herrera-Joancomartí and Cristina Pérez-Sola. Online social honeynets: Trapping web crawlers in osn. In *MDAI*, volume 6820 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 2011.

[71] José María Gómez Hidalgo. Evaluating cost-sensitive unsolicited bulk email categorization. pages 615–620, 2002.

[72] Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*, NAACL-Short '03, pages 34–36, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[73] Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20(1):19–62, May 2005.

[74] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[75] Alison Huettner and Pero Subasic. Fuzzy typing for document management. In *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, pages 26–27, 2000.

[76] Markus Jakobsson, Nathaniel Johnson, and Peter Finn. Why and how to perform fraud experiments. *IEEE Security and Privacy*, 6(2):66–68, 2008.

[77] Markus Jakobsson and Jacob Ratkiewicz. Designing ethical phishing experiments: a study of (ROT13) rOnl query features. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 513–522, New York, NY, USA, 2006. ACM.

[78] George H Jensen and John K DiTiberio. Personality and the teaching of composition, 1989.

[79] Nitin Jindal and Bing Liu. Mining comparative sentences and relations. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1331–1336. AAAI Press, 2006.

[80] Nitin Jindal and Bing Liu. Review spam detection. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1189–1190, New York, NY, USA, 2007. ACM.

[81] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 219–230, New York, NY, USA, 2008. ACM.

[82] Nitin Jindal, Bing Liu, and Ee-Peng Lim. Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1549–1552, New York, NY, USA, 2010. ACM.

[83] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, DTIC Document, 1996.

[84] Chris Kanich, Christian Kreibich, Kirill Levchenko, Brandon Enright, Geoffrey M. Voelker, Vern Paxson, and Stefan Savage. Spamalytics: an empirical analysis of spam marketing conversion. In *Proceedings of the 15th ACM conference on Computer and communications security*, CCS '08, pages 3–14, New York, NY, USA, 2008. ACM.

[85] Mark Kantrowitz. Method and apparatus for analyzing affect and emotion in text. U.S. Patent 6622140, 2003. Patent filed in November 2000.

[86] KasperskyLab. Spam in q3 2013: spike in malicious spam targeting user data. `http://www.kaspersky.com/about/news/spam/2013/Spam_in_Q3_2013_spike_in_malicious_spam_targeting_user_data`, november 2013.

[87] KasperskyLab. Spam and phishing in q1 2016. `https://securelist.com/analysis/quarterly-spam-reports/74682/spam-and-phishing-in-q1-2016/`, 2016.

[88] Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 32–38, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

[89] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[90] Yves Kodratoff. Knowledge discovery in texts: A definition, and applications. In *Proc. of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS-99*, pages 16–29. Springer, 1999.

[91] R Kishore Kumar, G Poonkuzhali, and P Sudhakar. Comparative study on email spam classifier using data mining techniques. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, pages 14–16, 2012.

[92] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[93] Raymond Y. K. Lau, S. Y. Liao, Ron Chi-Wai Kwok, Kaiquan Xu, Yunqing Xia, and Yuefeng Li. Text mining and probabilistic language modeling for online review spam detection. *ACM Trans. Manage. Inf. Syst.*, 2(4):25:1–25:30, January 2012.

[94] David D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Machine Learning: ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Proceedings*, pages 4–15, 1998.

[95] Y. H. Li and A. K. Jain. Classification of text documents. *The Computer Journal*, 41(8):537–546, 1998.

[96] E.D. Liddy. Natural language processing, NY. Marcel Decker, Inc 2001.

[97] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 939–948, New York, NY, USA, 2010. ACM.

[98] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. *Mining Text Data*, pages 415–463, 2012.

[99] Hugo Liu, Henry Lieberman, and Ted Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, pages 125–132, New York, NY, USA, 2003. ACM.

[100] Shah Mahmood and Yvo Desmedt. Online social networks, a criminals multipurpose toolbox (poster abstract). In Davide Balzarotti, Salvatore J. Stolfo, and Marco Cova, editors, *RAID*, volume 7462 of *Lecture Notes in Computer Science*, pages 374–375. Springer, 2012.

[101] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.*, 30(1):457–500, November 2007.

[102] Andreas Makridakis, Elias Athanasopoulos, Spiros Antonatos, Demetres Antoniades, Sotiris Ioannidis, and Evangelos P Markatos. Designing malicious

applications in social networks. *In IEEE Network Special Issue on Online Social Networks*, 2010.

[103] R Malarvizhi and K. Saraswathi. Content-based spam filtering and detection algorithms-an efficient analysis & comparison 1. *International Journal of Engineering Trends and Technology*, Vol. 4, Issue 9, September 2013.

[104] Shervin Malmasi and Mark Dras. Multilingual native language identification. *Natural Language Engineering*, FirstView:1–53, 12 2015.

[105] Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization*, 752:41–48, 1998.

[106] J.R. Méndez, F. Fdez-Riverola, F. Díaz, and J.M. Corchado. Sistemas inteligentes para la detección y filtrado de correo spam: una revisión. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 34:63–81, 2007.

[107] T. Mitchell. *Machine Learning*, pages 154–200. McGraw-Hill, 1997.

[108] Arjun Mukherjee, Bing Liu, Junhui Wang, Natalie Glance, and Nitin Jindal. Detecting group review spam. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 93–94, New York, NY, USA, 2011. ACM.

[109] Naresh Kumar Nagwani and Aakanksha Sharaff. Sms spam filtering and thread identification using bi-level text classification and clustering techniques. *Journal of Information Science*, pages 1–13, 2015.

[110] Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. 2013.

[111] Akshay Narayan and Prateek Saxena. The curse of 140 characters: evaluating the efficacy of sms spam detection on android. In *Proceedings of the Third ACM workshop on Security and privacy in smartphones & mobile devices*, pages 33–42. ACM, 2013.

[112] Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas By. Sentiment analysis on social media. In *ASONAM*, pages 919–926. IEEE Computer Society, 2012.

[113] M. T. Nuruzzaman, C. Lee, and D. Choi. Independent and personal sms spam filtering. In *Computer and Information Technology (CIT), 2011 IEEE 11th International Conference on*, pages 429–435, Aug 2011.

[114] Jon Oberlander and Scott Nowson. Whose thumb is it anyway?: Classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 627–634, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[115] Derek O'Callaghan, Martin Harrigan, Joe Carthy, and Pádraig Cunningham. Network analysis of recurring youtube spam campaigns. *CoRR*, abs/1201.3783, 2012.

[116] Bruno Ohana and Brendan Tierney. Sentiment classification of reviews using sentiwordnet. In *9th. IT & T Conference*, page 13, 2009.

[117] F.J. Ortega, J.A. Troyano, F. Cruz, and F. Enrıquez. Detección de spam en la web mediante el análisis de texto y de grafos. *IV Jornadas TIMM Tratamiento de la Información Multilingüe y Multimodal 7 y 8 de abril de 2011*, page 13, 2011.

[118] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 2004.

[119] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[120] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[121] Patrick Pantel and Dekang Lin. Spamcop: A spam classification & organization program. In *In Learning for Text Categorization: Papers from the 1998 Workshop*, pages 95–98, 1998.

[122] Constantinos Patsakis, Alexandros Asthenidis, and Abraham Chatzidimitriou. Social networks as an attack platform: Facebook case study. In Robert Bestak, Laurent George 0002, Vladimir S. Zaborovsky, and Cosmin Dini, editors, *ICN*, pages 245–247. IEEE Computer Society, 2009.

[123] Iasonas Polakis, Georgios Kontaxis, Spiros Antonatos, Eleni Gessiou, Thanasis Petsas, and Evangelos P. Markatos. Using social networks to harvest email addresses. In *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*, WPES '10, pages 11–20, New York, NY, USA, 2010. ACM.

[124] Iasonas Polakis, Marco Lancini, Georgios Kontaxis, Federico Maggi, Sotiris Ioannidis, Angelos D Keromytis, and Stefano Zanero. All your face are belong to us: breaking facebook's social authentication. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 399–408. ACM, 2012.

[125] Vipul Ved Prakash. Vipul's razor documentation. `http://www.kaspersky.com/about/news/spam/2013/Spam_in_Q3_2013_spike_in_malicious_spam_targeting_user_data`, 2013.

[126] Spamhaus Project. The spamhaus block list. `http://www.spamhaus.org/sbl/`, 2013.

[127] The Spamhaus Project. The definition of spam. `http://www.spamhaus.org/consumer/definition/`, 2013.

[128] Jefferson Provost. NaÃ¯ve-bayes vs. rule-learning in classification of email. Technical report, Department of Computer Sciences, University of Texas, Austin, 1999.

[129] Md Sazzadur Rahman, Ting-Kai Huang, Harsha V Madhyastha, and Michalis Faloutsos. Efficient and scalable socware detection in online social networks. In *USENIX Security*, 2012.

[130] Vallikannu Ramanathan and T. Meyyappan. Survey of text mining. In *International Conference on Technology and Business Management*, March 2013.

[131] Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter 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.

[132] Andreas Rauber and Alexander Müller-Kögler. Integrating automatic genre analysis into digital libraries. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, JCDL '01, pages 1–10, New York, NY, USA, 2001. ACM.

[133] Isidore Rigoutsos and Tien Huynh. Chung-kwei: a pattern-discovery-based system for the automatic identification of unsolicited e-mail messages (spam). In *CEAS*, 2004.

[134] Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe. Feature subsumption for opinion analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 440–448, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[135] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 105–112, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[136] Nazirova Saadat. Survey on spam filtering techniques. *Communications and Network*, 2011, 2011.

[137] Warren Sack. On the computation of point of view. In *Proceedings of AAAI*, page 1488, 1994. Student abstract.

[138] Georgios Sakkis, Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Constantine D. Spyropoulos, and Panagiotis Stamatopoulos. Stacking classifiers for anti-spam filtering of e-mail. *CoRR*, cs.CL/0106040, 2001.

[139] G. Salton and M. Smith. On the application of syntactic methodologies in automatic text analysis. *SIGIR Forum*, 23(SI):137–150, May 1989.

[140] Enrique Puertas Sanz, José María Gómez Hidalgo, and José Carlos Cortizo. Email spam filtering. *Advances in Computers*, pages 45–114, 2008.

[141] Santoshkumar Biradar Savita Teli. Effective spam detection method for email. In *International Conference on Advances in Engineering & Technology*, 2014.

[142] Jianqiang Shen, Oliver Brdiczka, and Juan Liu. Understanding email writers: Personality prediction from email messages. In *User Modeling, Adaptation, and Personalization*, pages 318–330. Springer, 2013.

[143] E. Simoudis. Reality check for data mining. *IEEE Expert*, 11(5):26–33, 1996.

[144] Jonghyuk Song, Sangho Lee, and Jong Kim. Spam filtering in twitter using sender-receiver relationship. In *Recent Advances in Intrusion Detection*, pages 301–317. Springer, 2011.

[145] Tao Stein, Erdong Chen, and Karan Mangla. Facebook immune system. In *SNS*, page 8. ACM, 2011.

[146] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 1–9, New York, NY, USA, 2010. ACM.

[147] P. Subasic and A. Huettner. Affect analysis of text using fuzzy semantic typing. *Fuzzy Systems, IEEE Transactions on*, 9(4):483–496, 2001.

[148] A. Tan. Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD Workshop on Knowledge Disocovery from Advanced Databases*, Beijing, 1999.

[149] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1397–1405, New York, NY, USA, 2011. ACM.

[150] Kenji Tateishi, Yoshihide Ishiguro, and Toshikazu Fukushima. Opinion information retrieval from the Internet. *Information Processing Society of Japan (IPSJ) SIG Notes*, 2001(69(20010716)):75–82, 2001.

[151] Brad Templeton. Reaction to the dec spam of 1978. `http://www.templetons.com/brad/spamreact.html`, 2008.

[152] Richard Tong. An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the SIGIR Workshop on Operational Text Classification*, pages 1–6, New Orleans, Louisianna, 2001.

[153] Konstantin Tretyakov. Machine learning techniques in spam filtering. In *Data Mining Problem-oriented Seminar, MTAT*, volume 3, pages 60–79. Citeseer, 2004.

[154] Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

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

[156] Alex Hai Wang. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10. IEEE, 2010.

[157] De Wang, Danesh Irani, and Calton Pu. A social-spam detection framework. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, pages 46–54. ACM, 2011.

[158] Janyce Wiebe, Eric Breck, Christopher Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury. Recognizing and organizing opinions expressed in the world press. In *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, 2003.

[159] Janyce Wiebe and Rebecca Bruce. Probabilistic classifiers for tracking point of view. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 181–187, 1995.

[160] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Comput. Linguist.*, 30(3):277–308, September 2004.

[161] Janyce M. Wiebe. Identifying subjective characters in narrative. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 401–408, 1990.

[162] Janyce M. Wiebe. Tracking point of view in narrative. *Comput. Linguist.*, 20(2):233–287, June 1994.

[163] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[164] Janyce M. Wiebe and William J. Rapaport. A computational theory of perspective and reference in narrative. In *Proceedings of the 26th Annual Meeting on Association for Computational Linguistics*, ACL '88, pages 131–138, Stroudsburg, PA, USA, 1988. Association for Computational Linguistics.

[165] Yorick Wilks. Information extraction as a core language technology: What is IE? In *Proceedings of Lecture Notes in Computer Science, chapter In M-T. Pazienza (ed.), Information Extraction*, pages 14–18. Springer, 1997.

[166] Yorick Wilks and Janusz Bien. Beliefs, points of view, and multiple environments. *Cognitive Science*, 7(2):95–119, 1983.

[167] John Wilson. *Politically speaking: The pragmatic analysis of political language*. Basil Blackwell Oxford, 1990.

[168] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 71–80. ACM, 2012.

[169] Chao Yang, Robert Chandler Harkreader, and Guofei Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *Recent Advances in Intrusion Detection*, pages 318–337. Springer, 2011.

[170] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[171] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 129–136, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[172] Xianghan Zheng, Zhipeng Zeng, Zheyi Chen, Yuanlong Yu, and Chunming Rong. Detecting spammers on social networks. *Neurocomputing*, 159:27 – 34, 2015.