Knowledge-Based Systems 108 (2016) 25–32

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering

Tiago A. Almeida[a,*], Tiago P. Silva[a], Igor Santos[b], José M. Gómez Hidalgo[c]

[a] Department of Computer Science, Federal University of São Carlos (UFSCar), Sorocaba, São Paulo, 18052-780, Brazil
[b] Universidad de Deusto, DeustoTech – Computing (S3Lab), Avenida de las Universidades 24, Bilbao, Vizcaya, 48007, Spain
[c] Analytics Department, Pragsis, Manuel Tovar 49-53, Madrid, 28034, Spain

ARTICLE INFO

Article history:
Received 30 October 2015
Revised 25 April 2016
Accepted 6 May 2016
Available online 14 May 2016

Keywords:
Instant Messaging spam filtering
SMS spam filtering
SPIM
Text categorization
Natural language processing

ABSTRACT

The rapid popularization of smartphones has contributed to the growth of online Instant Messaging and SMS usage as an alternative way of communication. The increasing number of users, along with the trust they inherently have in their devices, makes such messages a propitious environment for spammers. In fact, reports clearly indicate that volume of spam over Instant Messaging and SMS is dramatically increasing year by year. It represents a challenging problem for traditional filtering methods nowadays, since such messages are usually fairly short and normally rife with slangs, idioms, symbols and acronyms that make even tokenization a difficult task. In this scenario, this paper proposes and then evaluates a method to normalize and expand original short and messy text messages in order to acquire better attributes and enhance the classification performance. The proposed text processing approach is based on lexicographic and semantic dictionaries along with state-of-the-art techniques for semantic analysis and context detection. This technique is used to normalize terms and create new attributes in order to change and expand original text samples aiming to alleviate factors that can degrade the algorithms performance, such as redundancies and inconsistencies. We have evaluated our approach with a public, real and non-encoded data-set along with several established machine learning methods. Our experiments were diligently designed to ensure statistically sound results which indicate that the proposed text processing techniques can in fact enhance Instant Messaging and SMS spam filtering.

1. Introduction

Short text messaging is the mean of communication for a huge number of people nowadays. In this context, online Instant Messaging (IM) and SMS are clearly the leading technologies. In fact, it is estimated that about 80 billion messages are sent a day considering just SMS, WhatsApp and Facebook Messenger.[1]

SMS has become a massive commercial industry since messaging still dominates mobile market non-voice revenues worldwide. According to a Portio Research report,[2] the worldwide mobile messaging revenue was over 128 billion dollars in 2011, and in 2016 the revenue is forecasted to be over 153 billion dollars. The same document indicates that, in 2011, more than 7.8 trillion SMS messages were sent over the world, while more than 9.5 trillion were disseminated just in 2014.

In the same way, the popularization of smartphones along with low cost Internet plans are leading online Instant Messaging applications to become the means of electronic communication most used in the world. To get an idea, WhatsApp recently claimed to have over 1 billion users. Proportionally, this means that one in seven people on the planet use the messaging app.[3] According to a report released by Facebook in April 2016, about 70% of WhatsApp users access the application daily and more than 42 billion messages are sent a day. Moreover, Facebook Messenger has about 900 million monthly active users responsible for sending around 18 billion messages a day.

The growth in short text messaging along with unlimited texting plans allows malicious messages barely costs nothing for the attackers. This, combined with the trust users inherently have in their mobile devices, makes it a propitious environment for attack.

* Corresponding author.
E-mail addresses: talmeida@ufscar.br (T.A. Almeida), tpsilva@acm.org (T.P. Silva), isantos@deusto.es (I. Santos), jmgomez@pragsis.com (J.M. Gómez Hidalgo).

[1] SMS, Messenger and WhatsApp process 80 billion messages a day. Available at http://goo.gl/HdWm1v.

[2] Mobile Messaging Futures 2012-2016. Available at http://goo.gl/Wfb01z.

[3] F8 – Facebook Developer Conference (April 2016), available at http://goo.gl/HdWm1v

http://dx.doi.org/10.1016/j.knosys.2016.05.001
0950-7051/

As a consequence, Instant Messaging applications and SMS are becoming the latest target of electronic junk mail.

SMS spam (also known as mobile phone spam) is any junk message delivered to a mobile phone as text messaging. This practice, which became very popular in some parts of Asia, is now spreading in Western countries.[4] Besides being annoying, SMS spam can also be expensive since some users must pay to receive messages. Moreover, there is a very limited availability of mobile phone spam-filtering software and another concern is that important legitimate messages such as those of an emergency nature could be blocked. Nonetheless, many providers offer their subscribers means for mitigating unsolicited SMS messages.

More recently, the volume of spam is also increasing in similar environments. There are several indications that online Instant Messaging apps are the next target. Such messages are also known as SPIM – SPam over Instant Messaging. For instance, there is some evidence of chain letters and hoax in WhatsApp. Panda Labs have reported some of the most popular hoax in WhatsApp in Spain, in 2015, like the one that promises new emoticons if you click and send the same hoax (spam) to ten friends.[5] Acording to the company AdaptativeMobile, there are also spam campaigns in UK targeting WhatsApp users with investment spam messages sent from US numbers to Europe, spam promoting fake luxury goods sent from Chinese numbers to users in Europe, and others.[6] To avoid these messages, Facebook Messenger has added a feature to report a message as spam[7] and Skype has been reported by users as spammy as well.[8]

In traditional e-mail spam problem, simple techniques as blacklisting are often used to complement the content-based spam filtering. These solutions block e-mails from certain senders, whereas whitelisting [31] delivers e-mail from specific senders to reduce the number of misclassified ham e-mails. DNS blacklisting is one particular solution that checks the host address against a list of networks or servers known to distribute spam [33,49]. However, in IM and SMS spam domain, it is very difficult to having access to such data mainly because the providers must preserve the confidential data of their customers.

While companies are facing many problems in dealing with texting spam problem, academic researchers in this field are also experiencing difficulties. One of the concerns is that established email spam filters have their performance seriously degraded when used to filter SPIM or mobile phone spam. This happens due to the small size of these messages. Furthermore, these messages are usually rife of slangs, symbols, emoticons and abbreviations that make even tokenization a difficult task.

Noise in text messages can appear in different ways. The following phrase is an example: *"Plz, call me bak asap... Ive gr8 news! :)"*. There are misspelled words *"Plz, bak, Ive, gr8"*, abbreviation *"asap"* and symbol *":)"*. In order to transcribe this phrase to a proper English grammar, a *Lingo* dictionary[9] would be needed along with a standard English dictionary, which associates each slang, symbol or abbreviation to a correct term. After a step of text normalization, the input phrase would be transcribed to *"Please, call me back as soon as possible... I have great news! :)"*.

In addition to noisy messages, there are other well-known problems such as ambiguous words in context (polysemy) and different words with the same meanings (synonymy), that can harm the performance of traditional machine learning techniques when applied to text categorization problems.

Both synonymy and polysemy can have their effect minimized by semantic indexing for word sense disambiguation [45,54]. Such approaches associate meanings extracted from dictionaries to words by finding similar terms given the context of a message. In general, the effectiveness of using such dictionaries relies in the quality of terms extracted from samples. However, common tools for natural language processing can be not suitable to deal with short texts, demanding proper tools for work in such a context [7,18,41].

Even after dealing with problems of polysemy and synonymy, resulting terms may not be enough to classify a message as spam or legitimate because original samples are usually very short. In such a context, some recent works recommend employing ontology models to analyze each term and find associated new terms (with the same meaning) in order to enrich original sample and acquire more features [36,43].

In this scenario, we have designed and evaluated a text preprocessing approach to automatically normalize and provide semantic information for noisy and short text samples in order to enhance IM and SMS spam filtering. Our hypothesis is that such processing can increase the semantic information and consequently improve learning and predictions quality. Although such a proposal was evaluated in the context of SMS spam due to the availability of data, we highlight that our technique can also be applied to deal with messages sent by online Instant Messaging apps, since they have the same text characteristics.

In order to make use of semantic information, we have designed a cascade process in which we first transcribe the original messages from its raw form into a more standardized English language, in order to allow further and more accurate text analysis. We then extract semantic relations from the lexical database BabelNet [44], and apply Word Sense Disambiguation [1], intending to make this information more accurate. Finally, we expand the original message content with the extracted information, and make use of this normalized and expanded text representation to follow a traditional machine learning approach over the messages content. According to our experiments and statistical tests, this preprocessing can improve spam filtering effectiveness.

The remainder of this paper is organized as follows: in Section 2, we briefly review the main areas of interest covered in this work. Section 3 describes the proposed expansion method. In Section 4, we describe the dataset, performance measures and main settings used in the experiments. Section 5 shows the achieved results and details the performed statistical analysis. Finally, in Section 6, we present the main conclusion and outlines for future work.

2. Related work

Our work is mainly related to three research areas:

1. The employment of natural language techniques for chat and social media lexical normalization [29];
2. Using of lexical databases and semantic dictionaries in text representation for classification [24]; and
3. The applications themselves, namely content-based Instant Messaging and SMS spam filtering [3,25,39,40].

Lexical normalization is the task of replacing lexical variants of standard words and expressions normally obfuscated in noisy texts to their canonical forms, in order to allow further processing at text processing tasks. For instance, terms like "goooood" and "b4" should be replaced for the standard English words "good" and "before", respectively.

[4] Cloudmark annual report. Available at http://goo.gl/5TFAMM.

[5] Las 5 estafas de WhatsApp más famosas de 2015. Available at http://goo.gl/LY9gm7.

[6] Spammers set their sights on WhatsApp that's that ruined then. Available at http://goo.gl/yyya7D.

[7] How do I report a message as spam? Available at https://goo.gl/9qjkIr.

[8] Spoofed message from contact. Available at http://goo.gl/fw5wl4.

[9] Lingo is an abbreviated language commonly used on mobile and Internet applications, such as SMS, *chats*, emails, blogs and social networks.