ISSN 2007-9737

PS I Love You: Privacy Aware Sentiment Classification

Hugo Alatrista-Salas[1], Hugo Cordero[2], Miguel Nunez-del-Prado[1]

[1] Universidad del Pacífico,
Peru

[2] Universidad Nacional Mayor de San Marcos,
Peru

{h.alatristas, m.nunezdelpradoc}@up.edu.pe, h.cordero@unmsa.edu.pe

Abstract. At first glance, one might think that people are aware of the availability of comments or posts on social networks. Therefore, one may believe that people do not share sensitive information on public social networks. Nonetheless, people's posts sometimes reveal susceptible information. These posts include mentions the use of drugs or alcohol, sexual preferences, intimate confessions and even serious medical conditions like cancer or HIV. Such privacy leaks could cost someone to get fired or even worse to be a victim of denial insurance or bad credit evaluations. In this paper, we propose a complete process to perform a privacy-preserving sentiment analysis trough Bloom filters. Our approach shows an accuracy difference between 1% and 3% less than their classic sentiment analysis task counter part while guarantying a private aware analysis.

Keywords. Privacy, sentiment analysis, disclosure risk, information loss, bloom filter.

1 Introduction

Sentiment analysis attempts to determine whether an opinion, expressed in written text, is positive, negative or neutral, *w.r.t.*, entities or their characteristics [12]. These entities may be products, services, organisations, individuals, events or topics [22]. In a global survey on consumer sentiment[1], $10\ 000$ people were interviewed, and 93% of them did not feel comfortable with using information provided for purposes other than those for which it was collected. In addition, many companies do not take adequate measures to protect customer information and allow third parties to analyse it, even if data protection laws exist[2345]. This behavior discourages people from sharing data for fear of potential misuse, affecting the confidence in systems in which it is important to keep accurate information. We highlight this problem by way of the example of Sentiment Analysis task, which does not incorporate a mechanism to enhance user's privacy [1]. On the opposite, privacy-preserving Sentiment Analysis aims to get the polarity (positive or negative) without learning the underlying textual data.

In this context, we analyze three different corpora to predict the polarity of a document using the original versions and sanitized versions of them. Sanitization is perform by representing textual documents with Bloom Filters. Later, supervised learning algorithms were performed predict the polarity on the sanitized documents. Then, the trade-off between the knowledge acquired versus the privacy-preserving was measured through Information Loss and Disclosure Risk measures. Finally, our results were compared with results

[1] The Boston Consulting Group, "The Trust Advantage: How to win with Big Data", 2013

[2] Organisation for Economic Co-operation and Development (OECD), "OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data", 2013

[3] S. Cobb, "Data privacy and data protection: US law and legislation", 2016

[4] European Union, "General Data Protection Regulation (GDPR)", 2016

[5] Office of the Australian Information Commissioner, "Privacy (Credit Reporting) Code 2014 (version 1.2)", 2014

obtained using classic non-privacy aware methods of Sentiment Analysis to asses utility loss.

The rest of the document is organised as follows: Section 2 details the state of the art. Later, a complete description of privacy-preserving sentiment analysis is presented in Section 3. Thereafter, Section 4 gives details of the metrics used to evaluate the performance of our proposal. Then, the experiments are detailed in Section 5. This work ends with conclusions and future works.

2 Related Work

Concerning the sentiment analysis from textual documents via the use of machine learning methods, several studies were presented to the scientific community. For instance, in the work of Pang *et al.*[12], authors compare different algorithms, such as Naive Bayes, Support Vector Machines (SVM) and Maximum Entropy (ME). The authors obtained the best accuracy when using SVM.

In the same spirit, Vohra and Teraiya [19] uses two approaches: one based on a dictionary of words, and another one based on Machine Learning. The author shows that the second approach has better accuracy when improving the process to find words about the opinion in a particular domain. Concerning the second approach, the authors compared Naive Bayes, SVM and ME, in which the SVM algorithm got the best result (an accuracy of 95.55%).

Correspondingly, Singh and Husain [6] compared Naive Bayes, SVM, Multilayer Perceptron, and an unsupervised algorithm. The best accuracy was obtained by SVM (81.15%), *w.r.t.* other methods. Also, Boiy and Moens [3] examine SVM, Multinomial Naive Bayes MNB and ME over corpus in three different languages, namely English, Dutch and French achieving an accuracy of 83%, 70%, and 68%, respectively.

Xia *et al.* [21] use two types of feature sets: the part-of-speech based feature sets and the word-relation based feature sets. These two features are combined with three algorithms: Naive Bayes, ME and SVM algorithms. The authors conducted experiments using a document-level polarity corpus from Amazon[6], containing product reviews about four categories: books, DVDs, electronics, and kitchen. Authors obtained the best score (88.65%) using the SVM algorithm.

Other works use Logistic Regression for Sentiment Analysis [21]. For instance, the work of Mittal and Goel [11] uses Logistic Regression to analyse the correlation between market sentiment and public sentiment. The work uses data from Twitter to predict movements of the stock market. The authors obtained an accuracy up to 75.56%. Analogously, Thelwall *et al.* [17] uses Logistic Regression combined with SVM to attain an accuracy of 72.9%.

Methods above described, do not incorporate mechanisms to preserve data privacy because they only focus on the treatment and classification of textual data. Nevertheless, people express opinions on social networks, over time, revealing personal information explicitly or not, and in different circumstances [20, 10]. This kind of information can result in privacy leaks, which could become more dangerous if it is combined with external knowledge. For instance Humphreys *et al.* [5] analyse more than $2\,000$ tweets. This corpus was categorised in personal activities (66%), time (20.1%), proper names (22.7%), location (12.1%) and personal information (0.1%). In response to their findings, the authors suggest implementing privacy and protection mechanisms as social networks usage increases.

In the same spirit, Pang *et al.* [12] discuss the importance of data privacy and data manipulation in search engines based on opinion analysis. In particular, they highlight the privacy concerns raised by applications gathering data about peoples' preferences. The authors analysed different datasets ranging from public blogs to conversations. The challenge is to incorporate privacy mechanisms into the Sentiment Analysis task.

In this work, we rely on Bloom filters to achieve and guarantee some degree of privacy, while maintaining data utility. We have chosen Bloom Filter due to a different application in other privacy fields other than Text mining. For instance, record-binding applications [13], user's profiles

[6] Amazon: `www.amazon.com`

privacy mechanism [2], value masking in health domain [18], multi-biometric template-protected system [15], Chrome plug-in to gather Randomized Aggregatable Privacy-Preserving Ordinal Responses (RAPPOR) [4], encrypted search schemes [14]. It is worth noting that some papers use Bloom filter to improve speed and scalability [16, 7] but not privacy.

In the next section, the privacy-aware sentiment analysis process is detailed.

3 Privacy-aware Sentiment Analysis Process

The general scheme of our proposal is depicted in Figure 1. The mechanism takes a tagged corpus as an input. Therefore, the method starts with the bag-of-words construction through the term-frequency representation. For the term-frequency representation (M_{tf}) each word becomes part of the bag-of-words, whereby the frequency is computed as the number of occurrences of the word w_i for each document d_j belonging a corpus D. Depending on the context and the data, the matrix of terms representation can change to a matrix of term-frequency inverse-document-frequency (*tf-idf*), in which the tf measure is pondered by the number of occurrences of the term in the rest of the documents. In the present work, the term-frequency representation (tf) was arbitrarily used.

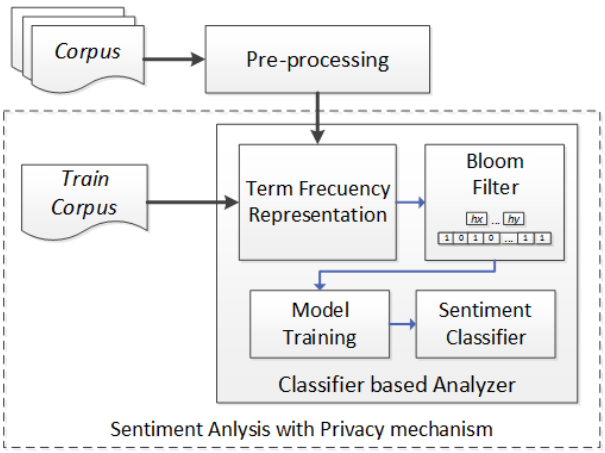

Fig. 1. Classifier-based scheme with privacy protection.

To provide privacy, we rely on Bloom filters before the training model task, *i.e.*, each word identified in the bag-of-words has a position in the Bloom filter defined by k hash functions. Thus, the hash functions generate the position for the words in the term-frequency matrix. When several hash functions are applied, each value obtained will be added to give the final position. To prevent an index out of bounds in the Bloom filter array, the module of each index value is calculated before being added. If the sum of all values exceeds the size of the array, a module is applied again. This process is parameterised by the number of hash functions k, the size of the array m and the number of words to add to the filter n. The combination of values of these three parameters could result in several words with the same index *i.e.*, words collisions. As a result, certain words would be lost in the vocabulary. Besides, documents $d_j \subseteq D$ would have fewer words w_i in the matrix of terms.

In this way, sanitization is introduced into the corpus D, through the words collision introduced by the hash functions of the Bloom filter. This new M'_{tf} matrix will be the input for the classification algorithm for training the model.

Once the model was generated on the sanitized corpus, measuring the effectiveness of our proposal is essential. This efficiency must be measured from two points of view: 1) efficiency of the classification task, using classic measures of machine learning such as accuracy, precision, among others; and, 2) how easy it is to re-identify a document after the Bloom filter has sanitised it. To measure the effectiveness of the Bloom filter, two measures were used in this work: the Disclosure Risk DR, and the Information Loss IL, whose are described in the next section.

4 Privacy Metrics

In the present paragraph, we describe the two metrics used to quantified the risk of an adversary re-linking a sanitized document with its original counter part and the amount of loss utility introduced by the privacy mechanism.

In broad terms, Disclosure Risk DR or the re-identification risk is the danger that a given form of disclosure will be encountered if a dataset is

released. In our work, the DR is estimated as the ratio between the number of correct re-identified documents and the total number of documents. Consequently, we take the similarity between the original and sanitised documents represented by the term frequency matrices M_{tf} and M'_{tf}. A couple of steps were performed to compute the DR value.

First, to compare the documents, we applied the transposed matrix of the original and sanitised frequency matrices ${M^t}_{tf}$ and ${M^{t'}}_{tf}$. Then, the Edit Distance ED is used as similarity measure [9]. The ED counts the number of operations required to transform one document into another (see Equation 1). In other words, the measure compares each position of the original document with its equivalent position in the sanitised document (*c.f.*, Equation 2), and if there are differences, the value of the Edit Distance is increased by one:

$$ed(d, d') = \sum_{i,j=0}^{n,m} diff(d[i][j], d'[i][j]), \tag{1}$$

$$diff(e, e') = \begin{cases} 1, \ if \ \ e \neq e', \\ 0, \ otherwise. \end{cases} \tag{2}$$

Then, the Edit Distance $ed(d, d')$ is the sum of differences between documents d and d'. To perform the re-link inference, a matrix M_{ed} should be constructed using Equation 1. The M_{ed} matrix size is $n \times n$, where n is the number of documents belonging the corpus D. Each position contains the value of the Edit Distance for each document dw compared to each sanitized document dw'.

Later, the Equation 3 is used to measure the distance between the original and sanitised document. This measure returns one (1) if the distance between both documents in the M_{ed} matrix is the minimum value. Finally, the ratio between the sum of coincidences of all rows in the matrix M_{ed} and the number of documents n is the Disclosure Risk DR (see Equation 4):

$$coinc(M_{ed}, i) = \begin{cases} 1, \ if \ M_{ed}[i, i]) = min(d_i), \\ 0, \ otherwise, \end{cases} \tag{3}$$

$$DR = \frac{\sum_{i=0}^{n} coinc(M_{ed}[i], i)}{n}. \tag{4}$$

Regarding Information Loss measure (IL), we rely on the False Positive Rate (FPR) to capture the proportion of the number non re-linked documents that are falsely count as re-linked. Equation 5 allows calculating the FPR. In this equation FP is the number of false positives, and TN is the number of true negatives:

$$FPR = \frac{FP}{FP + TN}. \tag{5}$$

The FPR allows understanding of the impact of the uncertainty applied to the corpus at our disposal. It is worth noting that both DR and IL metric are bounded. This allows a more suitable comparison when comparing different datasets.

5 Experiments

In the present section, we apply the privacy metrics to asses our model. For this purpose, we have used three different public available datasets [8]. For each dataset, there exists 500 positive and 500 negative sentences. The sentences of the datasets are from reviews of products[7], movies[8], and restaurants[9].

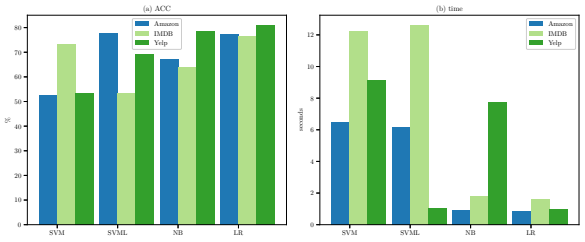

Fig. 2. Measure of a) performance and b) computation time of SVM with Radial basis function kernel, SVM with linear kernel, Naive Bayes and Logistic Regression without privacy mechanism.

To test our approach, we rely on four different classification algorithms, such SVM with

[7] Amazon: `amazon.com`
[8] Internet Movie Database (IMDb): `imdb.com`
[9] Yelp: `yelp.com`

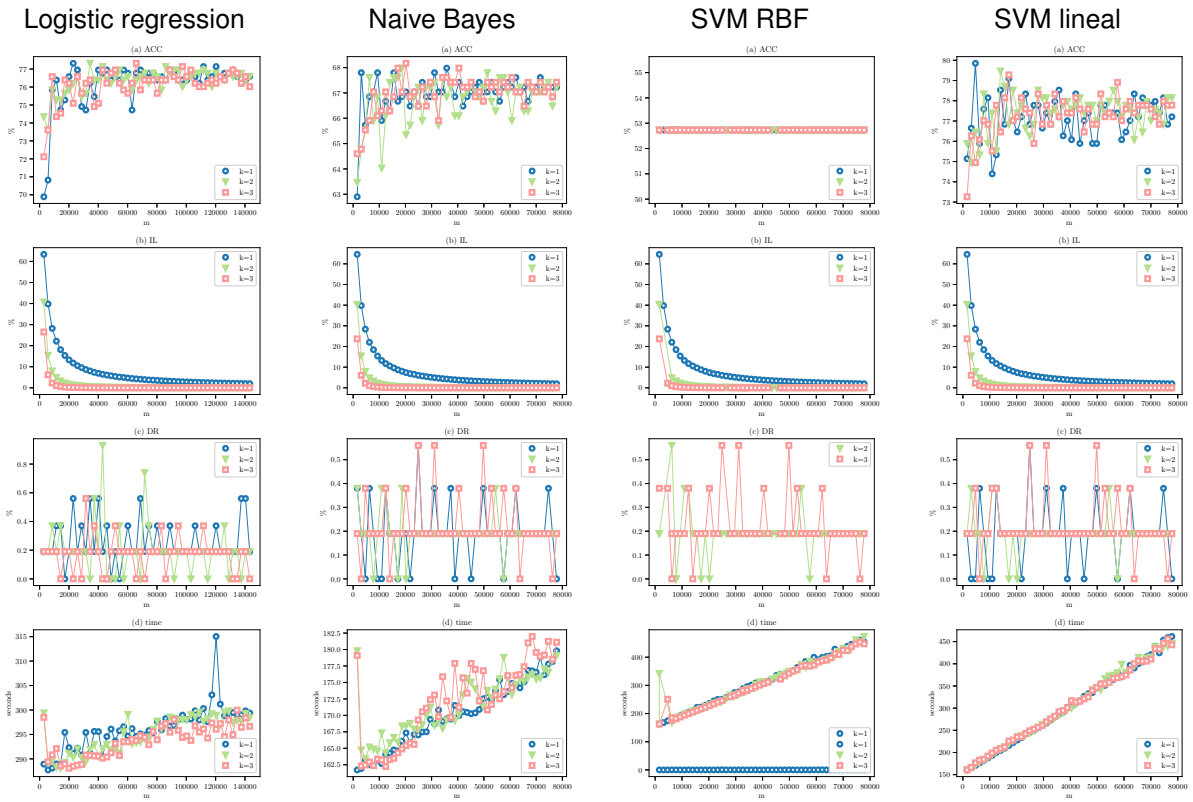

Fig. 3. Measure of a) accuracy (ACC), b) Information Loss (IL), c) Disclosure Risk and d) computation time (time) of the Logistic Regression, Naive Bayes, SVM with radial basis function kernel, and SVM with linear kernel with the privacy mechanism on the reviews of products dataset.

Radial basis function kernel, SVM with linear kernel, Naive Bayes and Logistic Regression. Therefore, we measure the accuracy and time of the four classification algorithms over the datasets mentioned above without the sanitisation mechanism.

Figure 2 depicts Logistic Regression as the best classification algorithm concerning accuracy and computation time for the three datasets. We also observe Naive Bayes and SVM with linear kernel performing as well as Logistic Regression for Yelp and Amazon datasets respectively,

To measure the Disclosure Risk (DR) and Information Loss (IL), we use the sanitization mechanism based on Bloom Filters over the reviews of products, movies, and restaurants datasets applying SVM with radial basis function kernel, SVM with linear kernel, Naive Bayes and Logistic Regression classification algorithms. It is worth noting that accuracy and time are also computed to measure the quantity of uncertainty and computational time introduced by the privacy mechanism.

Figure 3 shows different metrics to evaluate the performance of the sentiment classification while using the privacy mechanism. Thus, Logistic Regression obtains in average an accuracy of 76.33% with minimal and maximal values of 69.89% and 77.32%. Naive Bayes achieves in average 66.7034% ranging from 62.9% to 68.17%.

Concerning SVM RBF, the performance is 52.73%, which is constant even when the Bloom Filter size (m) changes. Finally, SVM with linear kernel attains in average 77.28% of accuracy with

ISSN 2007-9737

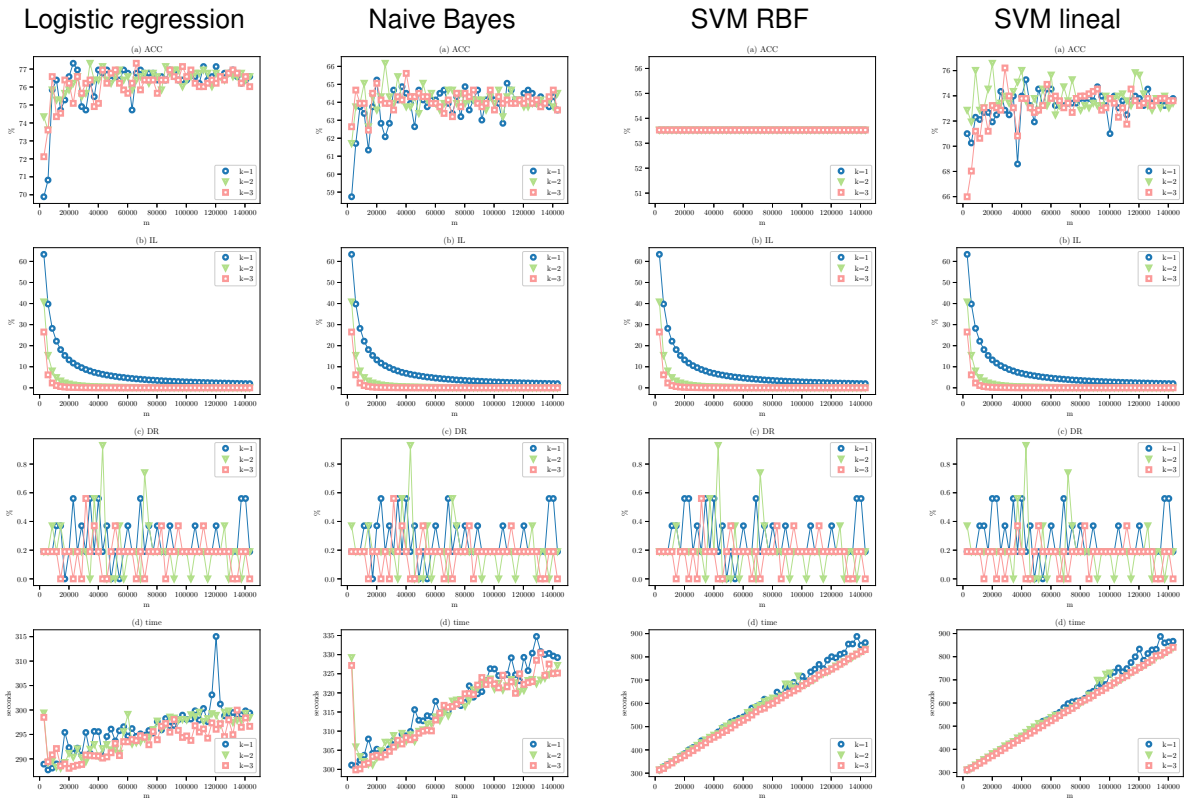

Fig. 4. Measure of a) accuracy (ACC), b) Information Loss (IL), c) Disclosure Risk and d) computation time (time) of the Logistic Regression, Naive Bayes, SVM with Radial basis function kernel, and SVM with linear kernel with the privacy mechanism on the movies dataset.

the minimum value of 74.39% and a maximum value of 79.47%. It is worth noting that the privacy mechanism does not degrade much the algorithms performance, but it adds more computational time as the Bloom Filter size increases independently of the number of Hash functions (k).

Regarding False Positive Rate and Information Loss decrease as the Bloom Filter size increments. This is an expected behavior since a bigger Bloom Filter avoids data collisions, which means less IL. On the other hand, Disclosure Risk (DR) values are very low ranging from 0 to 0.56 for the four algorithms. This value means that the privacy mechanism reduces the chances of an adversary to re-link comments.

Figure 4 shows the performance of the four classification algorithms on the movie dataset. Therefore, the average accuracy of Logistic Regression, Naive Bayes, SVN with RBF kernel and SVN with linear kernel are 76.33, 64.09, 53.53, and 73.84, respectively.

We observe that Logistic Regression does not loose accuracy as compared to the performance of its counterpart without the privacy mechanism. In addition, we note that the number of Hash functions does not impact significantly over accuracy.

On the contrary, time increases directly proportional to the size of the Bloom Filter. Concerning the IL,it decreases as the Bloom Filter size grows. We do not notice the influence of the number of Hash functions (k). Thus, the more Hash function we have, the fewer the probability of a false positive. Concerning DR, the probability of

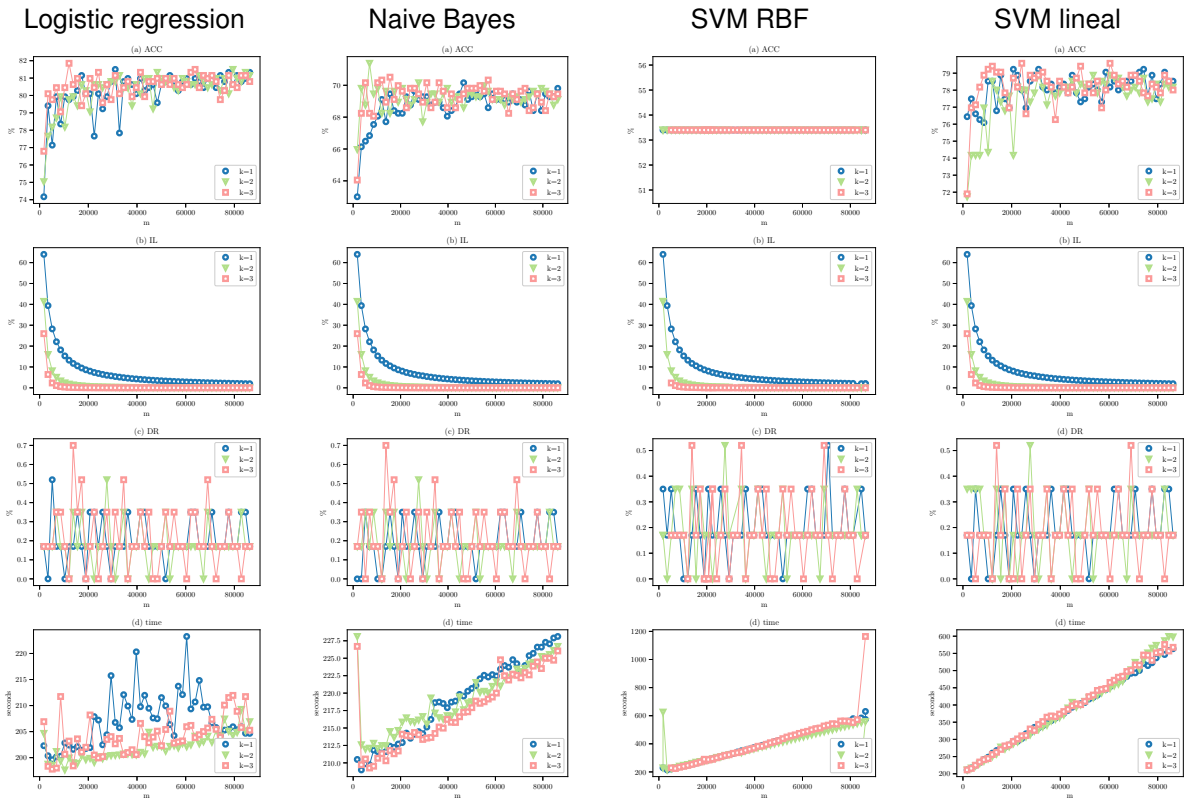

Fig. 5. Measure of a) accuracy (ACC), b) Information Loss (IL), c) Disclosure Risk and d) computation time (time) of the Logistic Regression, Naive Bayes, SVM with Radial basis function kernel, and SVM with linear kernel with the privacy mechanism on the restaurants dataset.

an adversary re-linking a message is between 0% and 0.93% with an average value of 0.27%.

The last experiment was using the restaurant dataset. Figure 5 reflects the performance of the classification algorithm when using the privacy mechanism. On the one hand, we notice that both Logistic Regression and SVM with linear kernel perform the best with average accuracy values of 80.26% a 78.093%, respectively. On the other hand, SVM with RBF kernel and Naive Bayes obtain an average accuracy of 53.4% and 69.24%, respectively. We observe that SVM with RBF kernel reminds constant despite the parameter configuration change, while Naive Bayes accuracy values are ranging from 63.0% to 71.38%.

Regarding computation time, we regard that time indirectly Proportional to the Bloom filter size. Apropos IL, as before observed, there is a decreasing trend while Bloom filter size grows, where the number of Hash function do not affects the IL. Finally, It is worth noting that DR value is in average 0.1822%. The minimal and maximal DR values are 0.0% and 0.7%, which makes difficult for an adversary to re-link textual comments.

6 Conclusions

Social networks allow mass spreading of opinions on a specific topic. Many of these opinions may contain sensitive information that may highlight uncomfortable situations for the person posting the message. In this context, text mining techniques - such as sentiment analysis - can leave this sensitive information uncovered. The idea is to

be able to perform the tasks of sentiment analysis preserving the personal information. To address this problem, privacy-aware text mining techniques appear as a solution to the problem of analyzing sensitive textual data.

In this article, we present a privacy-aware process to predict the polarity of textual documents using Bloom Filters. For this, our proposal was performed over three corpora, and we compared four supervised learning algorithms. Our results show that the logistic regression applied on Yelp corpus offers the best results (80% accuracy). On the contrary, SVM RBF offers the worst performance on Amazon corpus. Concerning the probability of an adversary re-linking a message, our results show an average value less than 0.30%. It means that out proposal guarantee the privacy of individuals. Is worth noting that, in all cases, the privacy mechanism does not degrade much the algorithms performance part while guarantying a private aware analysis.

As future works, we plan to implement a sentiment analysis process based on the use of dictionaries. The idea behind it is to add noise to the dictionaries in order to preserve privacy when predicting the polarity of the text. We also plan to analyse corpora that belong to more than two classes, for instance, corpus containing messages belonging a positive, negative or neutral connotation. In addition, we would like to asses our approach using bigger datasets. Finally, we want to try other supervised learning algorithms to obtain better results regarding the accuracy.

References

1. **Aggarwal, C. (2015).** *Data Mining: The Textbook*. Springer International Publishing.
2. **Alaggan, M., Gambs, S., & Kermarrec, A.-M. (2012).** BLIP: Non-interactive differentially-private similarity computation on Bloom filters. **Richa, A. W. & Scheideler, C.**, editors, *Stabilization, Safety, and Security of Distributed Systems*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 202–216.
3. **Boiy, E. & Moens, M.-F. (2009).** A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, Vol. 12, No. 5, pp. 526–558.
4. **Erlingsson, U., Pihur, V., & Korolova, A. (2014).** Rappor: Randomized aggregatable privacy-preserving ordinal response. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, ACM, New York, NY, USA, pp. 1054–1067.
5. **Humphreys, L., Gill, P., & Krishnamurthy, B. (2012).** How much is too much? privacy issues on twitter. *International Communication Association*.
6. **Husain, M. S. & Singh, P. k. (2014).** Methodological study of opinion mining and sentiment analysis techniques. *International Journal of Soft Computing (IJSC)*, Vol. 5, pp. 11–21.
7. **Kanavos, A., Nodarakis, N., Sioutas, S., Tsakalidis, A., Tsolis, D., & Tzimas, G. (2017).** Large scale implementations for twitter sentiment classification. *Algorithms*, Vol. 10, No. 1.
8. **Kotzias, D., Denil, M., de Freitas, N., & Smyth, P. (2015).** From group to individual labels using deep features. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, ACM, New York, NY, USA, pp. 597–606.
9. **Levenshtein, V. I. (1966).** Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, Vol. 10, pp. 707.
10. **Mao, H., Shuai, X., & Kapadia, A. (2011).** Loose tweets: An analysis of privacy leaks on twitter. *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*, WPES '11, ACM, New York, NY, USA, pp. 1–12.
11. **Mittal, A. & Goel, A. (2013).** Stock prediction using twitter sentiment analysis.
12. **Pang, B. & Lee, L. (2008).** Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, Vol. 2, No. 1-2, pp. 1–135.
13. **Schnell, R., Bachteler, T., & Reiher, J. (2009).** Privacy-preserving record linkage using Bloom filters. *BMC Medical Informatics and Decision Making*, Vol. 9, No. 1, pp. 41.
14. **Song, W., Wang, B., Wang, Q., Peng, Z., Lou, W., & Cui, Y. (2017).** A privacy-preserved full-text retrieval algorithm over encrypted data for cloud storage applications. *Journal of Parallel and Distributed Computing*, Vol. 99, pp. 14–27.
15. **Stokkenes, M., Ramachandra, R., Sigaard, M. K., Raja, K., Gomez-Barrero, M., & Busch, C. (2016).** Multi-biometric template protection — a security analysis of binarized statistical features for Bloom

filters on smartphones. *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6.

16. **Tayal, D. K. & Yadav, S. K. (2016).** Fast retrieval approach of sentimental analysis with implementation of Bloom filter on Hadoop. *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*, pp. 14–18.
17. **Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010).** Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, Vol. 61, No. 12, pp. 2544–2558.
18. **Vatsalan, D. & Christen, P. (2016).** Privacy-preserving matching of similar patients. *Journal of Biomedical Informatics*, Vol. 59, pp. 285–298.
19. **Vohra, S. M. & Teraiya, J. (2013).** A comparative study of sentiment analysis techniques. *Journal of Information, Knowledge and Research in Computer Engineering*, Vol. 02, No. 02.
20. **Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., & Cranor, L. F. (2011).** "i regretted the minute i pressed share": A qualitative study of regrets on facebook. *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, ACM, New York, NY, USA, pp. 10:1–10:16.
21. **Xia, R., Zong, C., & Li, S. (2011).** Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, Vol. 181, No. 6, pp. 1138–1152.
22. **Zhao, J., Liu, K., & Xu, L. (2016).** Sentiment analysis: Mining opinions, sentiments, and emotions. *Computational Linguistics*, Vol. 42, No. 3, pp. 595–598.

Article received on 27/12/2018; accepted on 05/03/2019.
Corresponding author is Hugo Alatrista-Salas.