Facebook Posts Text Classification to Improve Information Filtering

Randa Benkhelifa and Fatima Zohra Laallam
Department of Computer Science, Univ. Ouargla, BP 511 Route de Ghardaia, Ouargla 30 000, Algerie

Keywords: Facebook Posts, Text Classification, Pre-processing, Machine Learning Algorithms, Internet Slang.

Abstract: Facebook is one of the most used socials networking sites. It is more than a simple website, but a popular tool of communication. Social networking users communicate between them exchanging a several kinds of content including a free text, image and video. Today, the social media users have a special way to express themselves. They create a new language known as "internet slang", which crosses the same meaning using different lexical units. This unstructured text has its own specific characteristics, such as, massive, noisy and dynamic, while it requires novel preprocessing methods adapted to those characteristics in order to ease and make the process of the classification algorithms effective. Most of previous works about social media text classification eliminate Stopwords and classify posts based on their topic (e.g. politics, sport, art, etc). In this paper, we propose to classify them in a lower level into diverse pre-chosen classes using three machine learning algorithms SVM, Naïve Bayes and K-NN. To improve our classification, we propose a new preprocessing approach based on the Stopwords, Internet slang and other specific lexical units. Finally, we compared between all results for each classifier, then between classifiers results.

1 INTRODUCTION

Facebook is one of the largest social networking sites. Today, these are more than simple websites; they have become a very popular tool of communication between users. The latter can easily access and share content, news, opinions and information in general. This interactivity in the social media sites lands a new kind of text, which has specific characteristics, such as free, massive, dynamic, and noisy. Usually Internet users do not care about spelling and accurate grammatical construction of a sentence. They speak bravely using the slang terms (acronyms and abbreviations) in their posts. Sometimes users employ several lexical units that appear syntactically different but in fact they describe the same meaning. For example, ("$", "dlrs" or "dollars") represent the same thing "dollars", ("2$", "50$", "756$", "1dollars") represent "an amount of money", ("$", "dollars", "£", "¥") represent "a Currency", or in general all those represent "money". Generally, users create various text contents in the form of comments, wall posts, social media, and blogs. The majority of previous works classify a specific kind of posts sharing in social media in specific categories (e.g. news classification (Kovach, and Rosenstiel, 2007), (Nagar, 2009), (Weber, 2013), sentiment classification (Benkhelifa, and Laallam, 2015), (Liu, 2012), (Akaichi, et al., 2013), (Al-Ayyoub, et al., 2015) but they do not care about the topic of a random post. they do not cover some kind of posts such as "good morning friends". This cannot be into sport, economic or religion category. However, it is necessary to classify posts in a lower level in order to filter this kind of posts. Another point that we notice, is that most of research about text classification eliminates Stopwords. Words in a document that are frequently occurring but meaningless in terms of Information Retrieval (IR) are called Stopwords (Tsz-Wai, 2005). It is repeatedly claimed that Stopwords do not contribute towards the context or information of the documents and they should be removed during indexing as well as before querying by an Information Retrieval system (Tsz-Wai, 2005).. Two related facts were noticed in information retrieval in (Luhn, 1957). First of all, a relatively small number of words account for a very significant fraction of all text's size. Words like (IT, AND, MY and TO) can be found in virtually every sentence in English-based documents. Secondly, these words make very poor index terms (Amiri, and Chua, 2012). In this research, we will study the following questions: Are Stopwords always meaningless? Do they playing no role in improving classification regardless the

Benkhelifa, R. and Laallam, F.
Facebook Posts Text Classification to Improve Information Filtering.
In *Proceedings of the 12th International Conference on Web Information Systems and Technologies (WEBIST 2016) - Volume 1*, pages 202-207
ISBN: 978-989-758-186-1

categories?

Through the application of machine learning Algorithms, we propose to provide a classification of posts in categories that cover somehow the topics types discussed by users on the social network Facebook. The chosen categories are the following:

- News: to express neutral news usually originating from corporate Facebook pages (e.g. CNN, Fox news, etc) or personal news (e.g. i'm studying at new school now: the argan high school).
- Opinion: it conveys a positive or a negative sentiment, to express some feeling (e.g. I love mom), give an opinion (e.g. I think it was a great movie) or to ask for others' opinion (e.g. what do you guess about that?).
- Deal: it involves the patronage of various products (e.g. clothes) or services (e.g. hotel, restaurant).
- Demand: asking for something like job or money (e.g. I need a job. please anybody help me).
- Salutation: these are standard posts like "good morning" or "happy new year", etc.
- Quote: include all kind of quotes, proverbs and poem.

Text pre-processing plays a major role in any categorization. Despite the impotence of this phase, its implementation is too difficult especially due to the nature of the text generated from social networks. For the purpose of enhancing our classification, we propose a new pre-processing method with the benefit of the nature of the generated text from social networking. This method is based on: firstly preserving Stopwords, secondly, constructing a dictionary of top Internet slang terms, and another dictionary of the currencies mostly used in Facebook.

This article is organized as follows: In Section 2 we review the state of the art. Section3 describe the methodology, our proposed approach in Section 4. In Section 5, the obtained results are analyzed and discussed. Finally, our main conclusions and future work are drawn in Section 6.

2 RELATED WORK

Recently, several researches on text classification are crowned about text posts in social media.

In the researches (Vanetti, et al., 2013), (Uttarwar, and Bhute, 2013) the authors provide to the Online networking sites (ONS) users a system based on information filtering, which gives them a power to have a direct control to filter unwanted message (posts and comments) on their walls.

Most of research in social media have been addressed to tweets classification (Liu, 2012), (Al-Ayyoub, et al., 2015), (Sriram, et al., 2010), (Tang, et al., 2014), and also tweets clustering (Poomagal, et al., 2015).

Only a few studies have focused on Facebook posts classification. In (Akaichi, et al., 2013) performed sentiment classification on a novel collection of dataset which is Tunisian Facebook users' statuses. This collection leads to the Tunisian revolution, making a comparative experimental procedure between the Naïve Bayes and the SVM algorithms. (Faqeeh, et al., 2014) Focused on Short-text Document Classification in a cross-lingual setting: Arabic and English, and compared the performance of some of the most popular document classifiers on two datasets of short Facebook comments.

A few works in social media focus on internet slang. (Kundi, et al, 2014) Presented a framework on microblogs datasets for detection and scoring of Slang words for sentiment analysis. This framework enhances the pre-processing. (Amiri, and Chua, 2012) Focused on mining slang and urban opinion words and phrases from community-based question answering (cQA) archives.

3 METHODOLOGY

This section presents the methodology followed in this work. The objective of this work is to study the problem of short text special characteristics typically found on social media, and to show how much it is important to take these characteristics into consideration in the phase of pre-processing. Facebook posts are perfect for these due to their abundance and short length. Moreover, Facebook is a popular social network with a great diversity of users. This means that collecting a sufficiently large dataset with those characteristics on a various topics is feasible. To ensure the consistency and the reliability of our proposed approaches, we tested our classification and approaches on a collection of 20000 recent Facebook' text posts collected between (August and November 2015) from many Facebook profiles and pages. These texts were annotated manually by three human annotators to six different categories (4200 news, 6600 opinions, 3000 quotes, 3000 deals, 1500 salutations and 1700 demands).

3.1 Bags Development

This step focuses on the informal language used by

people on Facebook social networking. In this work, two types of bags were created: bag for Internet Slang Terms and bag for currency symbols.

We create a bag of the Top used Internet Slang Terms on Facebook, associated to their meaning. After a deep analysis, we concluded the results, which are more than 60 words. Some examples are shown in (Table 1).

Table 1: Examples of Internet slang terms.

Term	Meaning	Term	Meaning
2day	Today	THX	Thank you
2moro	Tomorrow	IRL	In Real Life
2nite	Tonight	ISO	In Search Of
BTW	By The Way	L8R	Later
B4N	Bye For Now	LOL	Laughing Out Loud
YA, U	You	NP	No Problem
UR	Your	N	And
GR8	Great	OMG	Oh My God
PLZ	Please	lil	Little

We created a bag of the currency symbols most used on Facebook social network (Table 2).

Table 2: Currency symbols.

Currency symbols			
$	$, dollars, dollar, dlrz, dlr, dlrs	¥	Yen, yuan
€	€, Euros, euro, eur,	£	Pound, pounds
¢	Cent, cents	₤	Lira, liras

3.2 Data Pre-processing

In this section, we show all algorithms developed and used to improving our classifications. Where Internet slang terms, money amounts or symbols and different values of percentage used by Facebook users are not really important. While their existence in the posts text is the important where the classifier takes each Internet slang term as a different lexical unit, But if we replace (internet slang terms by "SI", money's amount /symbols by "$", and the different values of percentage by "%", the classifier takes them as one same lexical unit, in other words, representing all symbols that express the same meaning by the same symbol to reduce the number of features and to make a stronger meaning in the text. To achieve this, we create the following algorithms:

a. Algorithm_1: detect and replace a set of Internet slang terms (dictionary in table 1) that are written in an irregular way by the right way (e.g. replacing U by you, GR8 by great).
b. Algorithm_2: detect and replace a set of Internet slang terms by the same lexical unit "SI".
c. Algorithm_3: Detect and replace all symbols or amounts of money by the same lexical unit "$", and all percentages by the lexical unit "%".
d. Pre-processing
   1. A term that appears less than 3 times is removed;
   2. Removing punctuation (.,!?) and symbols ([<>));
   3. The stemmer employed is the loven Stemmer which is used in the literature.
   4. Use TF-IDF (Ramos, 2003) as features' selection.

After the pre-processing and representing phase, various classifier algorithms can be applied. We have chosen three among the popular ones: K-Nearest Neighbor (K-NN) is one of the simplest methods for classification problem, Naïve Bayes (NB), and Multiclass Support Vector Machine (SVM). SVM is a binary classifier, to get M-class classifiers, construct set of binary classifiers, each trained to separate one class from rest.

4 PROPOSED APPROACH

In this section we describe alternative features token in each pre-processing method (P1-P6). Our proposed approach is as follows:

- P1: processing (removing stopwords).
- P2: processing (keeping stopwords).
- P3: algorithm_1, then the resulting document is processed (keeping stop words)
- P4: algorithm_2. Then the resulting document is processed (keeping stop words)
- P5: algorithm_3. Then the resulting document is processed (keeping stop words)
- P6: algorithm_2, then algorithm_3. The resulting document is processed (keeping stop words).

The document resulting from each version is classified each time by one classifier (K-NN, NB, and SVM) to predict the class of each post. Comparison is performed between all version results for each classifier. Then we compare between resulting classifiers.

5 RESULTS AND DISCUSSION

In this section we present the details of the experiments' concluded datasets and the analysis the

results providing some insights into the addressed problem. We started our experiment with analysing our datasets and showing which terms are the most frequent in each category from the six categories. The result is shown in Figure.1

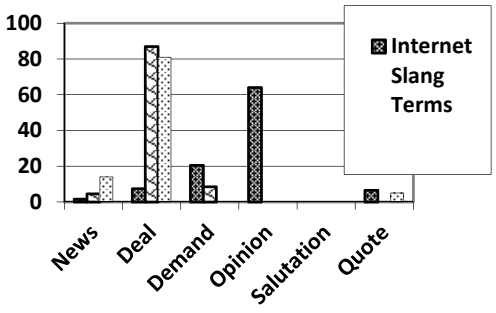

Figure 1: Distribution of terms in each class.

According to Figure.1, we note that Internet slang terms appear in deal, quote, demand, and for more than 60 % in opinion. This is related to the nature of this category: how someone's (simple user) speech is expressed using unformed language, contrasted with neutral news text, which is related more to express existence and facts by specialists using formal and correct language, where deal contains more than 85% of Money symbols/amounts and more than 80% of percentage. This is due to its nature which relates to business, buying, selling, prices and discounts. The rest is shared between news demand, and quote. Salutation contains none of those terms.

After this analysis and remarks, we decided to employ this characteristic in our pre-processing approaches in order to improve our classification.

The evaluation and the validation of our approaches are measured by three classifiers, which are: SVM, NB and KNN. SVM and NB classifiers are known to have high performance for the text classification (Aggarwal, and Zhai, 2012), and we aim to investigate whether this remains true for the social media text classification problem. And we added the KNN classifier because of its high sensitivity to the sparsity and its high dimensionality of the text classification problem in what is known as the "curse of the dimensionality" (Faqeeh, et al., 2014).

All the experiments were performed using WEKA tool (Hall, et al. 2009), where SVM, K-NN and Naïve Bayes is already implemented. Our choice is due to the added functionality it offers related to text pre-processing and analysis.

The performance of the classifiers under consideration and proposed pre-processing methods are measured by five widely used accuracy measures: precision (P), recall (R), F-measure (F-M), accuracy (A) and also the needed time (T) for execution.

For the data set validation, we chose using k-fold cross validation method with k=10, where data set is randomly divided into 10 equal sets "folds". The classification process is applied 10 times with one fold used for testing and the remaining nine folds training the classifier. This procedure is repeated for each of the 10 groups. All results are shown as follow:

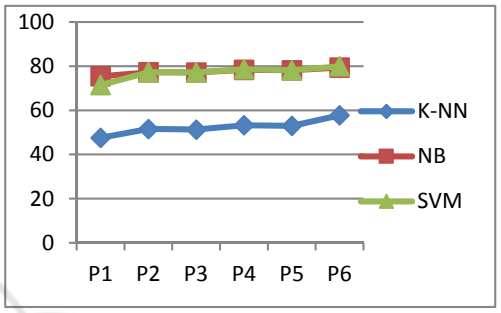

Figure 2: F-Measure results.

As shown in (figure 2), the best results by K-NN, NB and SVM classifiers is gotten by using P6 Pre-processing to our dataset. The detailed results are exposed in (table 3).

Table 3: Compare results.

Classifier		P (%)	R (%)	F-M (%)	A (%)	T (s)
KNN	P1	57.6	52.2	47.6	52.2	**0.01**
	P2	62.8	55.6	51.6	55.57	**0.01**
	P3	62.2	55.5	51.3	55.47	**0.01**
	P4	69.2	57.3	53.3	57.3	**0.01**
	P5	64.2	56.7	53	56.72	**0.01**
	P6	64.6	61.3	57.8	61.3	**0.01**
NB	P1	77.4	75.5	75.5	75.5	6.1
	P2	77.8	77.3	77.3	76.86	10.62
	P3	77.7	77.3	77.2	76.11	9.53
	P4	79.7	78.2	78.4	78.4	5.83
	P5	78.6	78.2	78.2	77.26	9.92
	P6	80.6	79.4	79.4	79.4	5.5
SVM	P1	72.9	71.7	71.5	71.7	8.28
	P2	77.8	77.3	77.3	77.31	15.98
	P3	77.7	77.3	77.2	77.26	15.16
	P4	79.6	78.9	78.7	**78.9**	12.01
	P5	78.6	78.2	78.2	**78.21**	13.8
	P6	80.9	80.2	80	**80.3**	10.15

On the scale of time, K-NN classifier is the fastest one among the three classifiers, but it is also the least performing in terms of accuracy. The highest accuracy we got is **80.3**%, by using SVM classifier and P6 pre-processing.

Firstly, we have removed Stopwords as their removal has a remarkable impact on decreasing the results; in contrast, keeping them enhances accuracy for K-NN classifier (from 52.2 to 55.57), NB (for 75.5 to 76.86), and SVM (from 71.7 to 77.31). The results decelerate the classification process. This indicates that they are sometimes very important and meaningful especially in this kind of classification. Using P3 pre-processing and comparing to P2 pre-processing process, there was a decrease of the accuracy for K-NN classifier (for 55.57% to 55.47%), NB (from 76.86% to 76.11%), and SVM (for 77.31% to 77.26%). This indicates that the presence of Internet Slang Terms enhances the performance of the classification. The results obtained by using P4 pre-processing confirm that the accuracy's performance is improved for K-NN classifier (for 55.57% to 57.3%), NB (from 76.86% to 78.4%), and SVM (for 77.31% to **78.9**%), and also in terms of Time for NB classifier (from 10.62s to 5.83s) and SVM (for 15.98s to 12.01s). Moreover, when we compare the P5 to the P2 pre-processing processes, there is a slight improvement for K-NN accuracy (from 55.57 to 56.72), NB (from 76.86 to 77.26), and SVM (from 77.31 to 78.21), and in time also for NB classifier (from 10.62s to 9.92s) and SVM (for 15.98 to 13.8s).

By gathering the two pre-processing P4 and P5 in P6, we got better performance of our three classifiers' accuracy (from 55.57% to 61.3%) for K-NN, (from 76.86% to 79.4%) for NB, and (from 77.31% to **80.3**%) for SVM, and also in time (from 10.62s to 5.5s) for NB and SVM (for 15.98 to 10.15s).

The three classifiers reached their highest accuracies when P6 pre-processing is applied on our dataset. However, the lowest accuracies were reached when we use a P1 Pre-processing.

Table 4 shows the best results obtained by SVM classifier for each pre-processing dataset version for each category.

Table 4: SVM results.

Category	Classifier	SVM			
	Measure	P	R	F-M	A (%)
News	P1	0.777	0.712	0.743	71.2
	P2	0.789	0.748	0.768	74.8
	P3	0.797	0.748	0.771	74.8
	P4	0.748	0.798	0.772	79.8
	P5	0.779	0.748	0.763	74.8
	P6	0.754	0.815	0.783	81.5
Deal	P1	0.867	0.763	0.812	76.3
	P2	0.897	0.817	0.855	81.7
	P3	0.881	0.813	0.846	81.3
	P4	0.859	0.797	0.827	79.7
	P5	0.928	0.857	0.891	85.7
	P6	0.931	0.791	0.855	79.1
Demand	P1	0.758	0.588	0.662	58.8
	P2	0.799	0.794	0.796	79.4
	P3	0.79	0.776	0.783	77.6
	P4	0.967	0.694	0.808	69.4
	P5	0.808	0.794	0.801	79.4
	P6	0.938	0.718	0.813	71.8
Opinion	P1	0.641	0.829	0.723	82.9
	P2	0.712	0.823	0.764	82.3
	P3	0.712	0.823	0.764	82.3
	P4	0.76	0.899	0.824	89.9
	P5	0.726	0.832	0.775	83.2
	P6	0.774	0.903	0.833	90.3
Salutation	P1	0.691	0.689	0.689	68.7
	P2	0.793	0.767	0.78	76.7
	P3	0.792	0.787	0.789	78.7
	P4	0.961	0.827	0.889	82.7
	P5	0.804	0.767	0.785	76.7
	P6	0.933	0.84	0.884	84
Quote	P1	0.722	0.52	0.605	52
	P2	0.77	0.647	0.703	64.7
	P3	0.776	0.647	0.705	64.7
	P4	0.698	0.564	0.624	56.4
	P5	0.767	0.647	0.702	64.7
	P6	0.703	0.601	0.648	60

By analyzing the results obtained by SVM classifier, we remark that replacing Internet slang terms by the same lexical unit in P4 improves the results especially in Opinion category by 7% in accuracy. And replacing all amounts/symbols of money by the same symbol and all percentages by the same symbol in P5 improves the result especially in deal category by 4% in accuracy.

6 CONCLUSION AND FUTURE

This paper is aimed at classifying Facebook text posts according to a new set of selected categories (i.e., news, opinion, deal, demand, salutation and quote). The main focus of this work is on the novel preprocessing method proposed and described therein, rather than on the selected classification techniques. Experiments have been made with three kinds of classifiers (i.e., K-NN, NB, and SVM), using 10-fold cross validation. A kind of comparative assessment has been made by comparing the results obtained from the selected algorithms with preprocessing. Interestingly enough, We conclude that stopwords are not always meaningless, as they play a major role in improving the performance of some classification, depending

on the category at hand. Also Internet slang terms, which contribute to improve the classification performance.

As a future extension of this work, we plan to explore other languages especially French and Arabic dialectal languages besides English. We will also include more categories, and propose other pre-processing approaches based on other features.

REFERENCES

Aggarwal, C. C., and Zhai, C. 2012. "A survey of text classification algorithms", In *Mining text data*, Springer US, pp. 163-222.

Akaichi, J., Dhouioui, Z., and Lopez-Huertas Perez, M. J. (2013) "Text mining facebook status updates for sentiment classification". In *System Theory, Control and Computing (ICSTCC), 17th International Conference, IEEE,* pp. 640-645.

Al-Ayyoub, M., Essa, S. B., & Alsmadi, I., 2015. " Lexicon-based sentiment analysis of Arabic tweets", *International Journal of Social Network Mining*, Vol.2, No.2, pp.101 – 114.

Amiri, H., and Chua, T. S. 2012. "Mining slang and urban opinion words and phrases from cQA services: an optimization approach". In *Proceedings of the fifth ACM international conference on Web search and data mining*, ACM, pp. 193-202.

Belew, R. K. 2000. *Finding out about: a cognitive perspective on search engine technology and the WWW*, Vol. 1. Cambridge University Press.

Benkhelifa, R., Laallam, F.Z, 2015. "Opinion Extraction and Classification of Real Time E-commerce Websites Reviews", *International Journal of Computer Science and Information Technologies*, Vol. 6 No. 6 , pp 4992-4996.

Faqeeh, M., Abdulla, N., Al-Ayyoub, M., Jararweh, Y., and Quwaider, M. 2014. "Cross-lingual short-text document classification for facebook comments". In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on. IEEE.* pp. 573-578.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. 2009. "Witten, The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, Vol. 11, No. 1.

Hu, X., Tang, J., Gao, H., and Liu, H. 2013. "Unsupervised sentiment analysis with emotional signals". In *Proceedings of the 22nd international conference on World Wide Web*, *International World Wide Web Conferences Steering Committee*, pp. 607-618.

Kovach, B. and Rosenstiel, T. 2007. "The Elements of Journalism: What Newspeople Should Know and the Public Should Expect". *Three Rivers Press*.

Kundi, F. M., Ahmad, S., Khan, A., and Asghar, M. Z. 2014. "Detection and Scoring of Internet Slangs for Sentiment Analysis Using SentiWordNet", *Life Science Journal*, Vol.*11* No. 9.

Liu, B. 2012. "Sentiment analysis and opinion mining". *Synthesis Lectures on Human Language Technologies*, Vol 5, No, 1 pp. 1–167.

Luhn, H. P., 1957. "A statistical approach to mechanized encoding and searching of literary information". *IBM Journal of Research and Development*, Vol. 1 No. 4, pp 309–317.

Nagar, N.a. 2009. "The Loud Public: Users' Comments and the Online News Media". *Online Journalism Symposium*.

Poomagal, S., Visalakshi, P., and Hamsapriya, T. 2015. "A novel method for clustering tweets in Twitter. *International Journal of Web Based Communities"*, Vol. *11 No.* 2, pp 170-187.

Ramos, J. 2003. "Using tf-idf to determine word relevance in document queries". In *Proceedings of the first instructional conference on machine learning*.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M., 2010. Short text classification in twitter to improve information filtering. *In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval,* ACM. pp. 841-842.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. 2014. "Learning sentiment-specific word embedding for twitter sentiment classification". In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Vol. 1, pp. 1555-1565.

Tsz-Wai Lo, R., He, B., and Ounis, I. 2005, "Automatically building a stopword list for an information retrieval system". In *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, Vol 5, pp 17–24.

Uttarwar, M., and Bhute, Y., 2013. "A Review on Customizable Content-Based Message Filtering from OSN User Wall" *IJCSMC*, Vol. 2, No. 10, pp 198 – 202.

Vanetti, M., Binaghi, E., Ferrari, E., Carminati, B., and Carullo, M. 2013. "A System to Filter Unwanted Messages from OSN User Walls", *IEEE Trans. Knowledge and Data Eng.*, Vol. 25, No. 2, pp. 1041-4347.

Weber, P. 2013. "Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments". *New Media & Society*, Vol.*16 No.* 6, pp 941-957.