Classification of Polish Email Messages: Experiments with Various Data Representations

Jerzy Stefanowski and Marcin Zienkowicz

Institute of Computing Science, Poznań University of Technology,
ul. Piotrowo 2, 60-965 Poznań, Poland
Jerzy.Stefanowski@cs.put.poznan.pl, marcas@op.pl

Abstract. Machine classification of Polish language emails into user-specific folders is considered. We experimentally evaluate the impact of different approaches to construct data representation of emails on the accuracy of classifiers. Our results show that language processing techniques have smaller influence than an appropriate selection of features, in particular ones coming from the email header or its attachments.

1 Introduction

An automatic categorization of emails into multiple folders could help users in filtering too many incoming emails and organizing them in a structure corresponding to different user's topics of interest. We are interested in applying machine learning methods to find the user's folder assignment rules based on the examples of his previous decisions [2, 4]. Let us remark the *folder categorization* is a more general problem than an identification of spams only [1]. It is also different to traditional text categorization as email messages are poorly structured comparing to longer texts and are written in an informal way. Moreover, there is no standard way of preparing email representation and there have been no multiple folder benchmark data sets; a recent exception is the Enron corpus [3].

Because of the page limit we reduce discussion on related works, see e.g. reviews in [1, 3]. Shortly speaking, the current research are focused on evaluating accuracy of various classifiers created by learning approaches with indication to Naïve Bayes, support vector machines, k-nearest neighbor or boosted decision trees. Most of these studies concerned spam filtering than folder categorization.

Unlike these research we focus our attention on issues of constructing proper data representation, i.e. transforming emails into a data format suitable for learning algorithms. The most commonly used is the *bag of words* representation. Documents are represented as vectors of feature values referring to the importance of words selected from messages. The features are mainly extracted from text parts of an email *subject* or its *body*. Some researchers use also other features acquired from the email header and from *attachment* files [2–4]. However, the role of these features in obtaining efficient classifiers is not enough discussed.

Moreover, nearly all research concerned English emails. There are no such works for the Polish language, which has more complex inflection and less strict

word order. Our previous study on Web page clustering showed that preprocessing of documents in Polish influenced the results more than in English [5].

Therefore, the aim of our paper is to experimentally study the influence of different ways of constructing data representation on the accuracy of typical classifiers in the task of categorization Polish language emails into user-specific folders. As there is no corpora of real email messages written in Polish we have first to collect them. In data representations, we want to explore the use of additional information available from the header and attachment parts, which were not commonly applied in the previous research. Moreover, we will examine the different ways of handling Polish language stemming and word extraction.

2 Data Sets and Their Representation

We collected three different Polish language email data sets. The first one, further called *MarCas*, is a copy of personal emails received by the second author of this paper since October 2003 till October 2004 and contained 4667 messages. Nine different categories (folders) were used within this period (the number of emails is reported in brackets): *BooBoo* – name of the student group (218 messages), *Humor* – jokes, stories (259), *pJUG* – Java group (179), *Work* (1119), *Family* (232), *Study* (587), *Friends* (289), *Spam* (1521), *Other* (263). The other set, called *DWeiss*, is coming from our close collaborator Dawid Weiss and contains the copy of his mailbox from a longer period: October 2001-2004. Unfortunately, it concerns spam detection only and contains 9725 messages from two categories legitimate *mail* – 5606 and *spam* – 4119. Both collections contained complete information about header sections and all attachments.

To extend information about multiple folders we constructed a third data set as a collection of 36260 messages from newsgroups in Polish recorded since March till April 2005. The *NewsGroups* contains messages from the following groups: pl.comp.lang.java (2971), pl.comp.lang.php (4896), pl.rec.kuchnia – Polish cuisine (4137), pl.rec.muzyka.gitara – guitar music (7844), pl.regionalne.poznan – regional news about the city of Poznan (4896), pl.regionalne.poznan – news about Warsaw (6604), pl.sci.fizyka – physics (2566), pl.sci.metamatyka – mathematics (2286). Unfortunately, available information on messages is restricted, i.e. no attachments were stored in the NewsGroups and header parts were somehow restricted to main sections only, e.g. a section on message routing was missing.

Let us discuss a construction of data representations. The text content of the *email body* part was the basis for extracting features for the bag of words representation. The *header* part includes the set of pairs (*parameter-value*) – their meaning is defined in RFC standards. Similar to some researchers, we parsed the file to identify the following elements: *Subject*, sender information (*From, Replay to*) and recipient (*To, CC, Bcc*). The subject field was handled in the same way as the body. Other elements were parsed to get the complete email addresses or nicknames. We also distinguished the case of multiple recipients, i.e. each led to a separate feature. The recipient numbers were additional numeric features. Unlike previous researchers we constructed next features based on information about

Table 1. Cardinality of features of given types in studied data

data	header	servers	subject	body	attach
MarCas	99	54	15	160	75
DWeiss	42	28	11	137	283
NewsGroups	24	0	12	198	0

other parameters stored in the header. Quite important for further analysis were *Received* parameters describing the route of the message by computer servers. We also defined a numeric attribute representing the number of servers, which were recorded for the given message. Additional features included the message encoding, tools for handling it, date, parts of the day, size of the message.

Besides these extended header features, we used more features about attachments, i.e. numeric ones representing the number of attached files of a given type (zip, jpg, exe), names of the most frequent files. Contents of non-binary attachment files were processed in the same way as the body or the subject and their terms were used in the final representation.

While processing text parts we encountered the problems of *stemming*, *word extraction* and handling extra *formatting tags* in HTML. Although previous researchers were not consistent on their role in English emails [3, 2, 4], the proper use of stop words and stemming is more influential on processing Polish documents [5]. To examine it in email classification we created different versions of each data set: either with or without stemming. We used a quasi-stemmer for Polish called *Stempelator* introduced by Weiss the text mining framework *Carrot*^{2}; for more details see `www.carrot2.org`. Similarly we studied two versions of word extraction: (1) with white characters tokenization only, (2) also with handling the additional characters as "@", "$" or "!". As to email bodies written in the HTML format we also tested two techniques: either these tags were removed or not – their names were used as additional parameters. To sum up, depending on the above choices we had 8 versions of the data sets.

Because of page limits we skip details of all carried out experiments with creating classifiers from these data sets and present conclusions only. For all data sets tokenization should be extended by extra characters. Stemming was particularly useful for DWeiss data while being not so significant for others. HTML tags should be removed from MarCas and NewsGroups and should stay in DWeiss data (perhaps because of their specific use in spam).

As the number of features in data representations was quite high (around few thousands) we decided to reduce their number. Firstly, while creating bag of words representation, we removed too frequent and too infrequent terms. Then, we employed a typical feature selection mechanism based on the following measures: *information gain*, *gain ratio* and χ^2 statistics. The single features were ranked according to these measures. For each data set, we selected the features occurring in the first positions over mean values of the evaluation measures in these rankings. The general characteristics of reduced data is given in Table 1,

Table 2. Total accuracy (%) for different versions of data representations

data set	learning algorithm	feature set subject	body	all
MarCas	IB3	61.0	70.3	88.3
	J48	61.4	70.2	87.7
	Naïve Bayes	59.2	48.3	80.0
DWeiss	IB3	82.9	83.5	98.6
	J48	82.7	82.6	98.4
	Naïve Bayes	82.3	79.3	82.2
NewsGroups	IB3	32.7	57.3	59.3
	J48	32.6	66.4	70.6
	Naïve Bayes	30.0	51.9	66.7

where we distinguished the following types of features depending on their origin: body, subject, servers, attachments, header (additional features created from other parts of the header, in particular senders).

3 Experiments

Three learning algorithms were chosen to construct email classifiers: k-nearest neighbor (k-NN), decision tree and Naïve Bayes. For running our experiments we used the Weka toolkit[1]. J48 (Weka C4.5 decision tree implementation) was used with standard options. IBk – a k-NN algorithm implementation was tested with the number of neighbors equal to 1, 3 and 5 and weighted Euclidean distances. The best results were obtained for $k = 3$. In Naïve Bayes we used an option of discretizing numerical attributes.

Performance of classifiers was evaluated with the following measures: *total accuracy*, *recall*, *precision* and *F-measure*. In our experiments we used *F-measure* with the aggregation function which assigned equal importance to both precision and recall. All these measures were calculated for each category/ folder. Stratified 10-fold cross validation was applied for all data sets and we report averaged results. They are presented in Tables 2 and 3. Because of page limits, Table 3 contains results for 3-NN (IB3) only, as it performed similarly to J48 and they both were better than Naïve Bayes.

4 Discussion of Results and Conclusions

Let us discuss results of experiments. Comparing learning algorithms we noticed that Naïve Bayes produced significantly worse classification accuracy than other algorithms for all versions of MarCas and DWeiss data sets – the accuracy was at least 10% lower. For NewsGroups the difference of accuracy was not so large. Looking for the best classifiers, we could say that for e-mail data MarCas and

[1] see `www.cs.waikato.ac.nz/ml/weka`

Table 3. Performance of classifiers for separate folders. For each measure, three successive numbers [in %] refer to three types of features sets: subject, body and all features.

data set	folder	Recall			Precision			F-measure		
MarCas	BooBoo	84.4	70.9	86.9	97.2	97.2	97.2	88.9	69.9	87.9
	Humor	13.7	9.9	70.8	60.4	44.2	71.4	22.1	15.8	71.1
	pJuG	93.4	57.2	96.4	95.8	96.2	97.2	94.3	72.2	96.6
	Work	82.2	73.5	90.3	41.8	75.5	90.1	55.4	72.8	91.3
	Family	29.2	72.8	91.8	72.2	90.6	93.1	41.7	83.0	92.6
	Study	35.7	53.9	83.5	87.9	67.5	85.4	50.8	59.7	85.4
	Friends	11.3	30.4	49.5	80.0	58.7	53.7	19.5	40.1	49.6
	Spam	76.5	96.2	97.6	74.0	63.2	95.6	75.2	77.3	98.1
	Other	28.5	66.1	71.9	66.3	80.4	77.5	36.8	72.8	74.6
DWeiss	mail	85.1	79.3	98.4	81.3	98.3	98.7	86.0	83.1	98.8
	spam	76.3	80.2	97.8	82.4	72.2	97.9	77.6	83.2	98.4
NewsGroups	java	9.3	65.6	66.5	44.5	67.2	64.6	10.5	51.0	44.5
	php	38.6	78.7	80.0	47.0	78.3	78.5	38.6	63.3	64.4
	kuchnia	22.6	62.2	64.8	43.6	62.4	63.7	23.7	55.0	55.3
	gitara	40.4	70.7	76.1	28.4	65.7	75.3	40.3	62.3	65.8
	poznan	31.1	68.3	72.6	43.8	70.8	73.8	31.2	60.4	64.4
	warszawa	29.4	60.8	66.0	27.8	59.3	64.8	30.1	50.8	55.7
	fizyka	31.6	57.0	63.2	51.4	61.1	67.5	31.8	54.0	54.7
	matematyka	9.1	61.6	68.4	23.4	65.0	71.4	9.4	56.6	58.1

DWeiss the performance of decision tree and kNN algorithms was comparable, while J48 tree algorithm was slightly better for NewsGroups. The problem of detecting spam was easier task than multiple folder categorization.

More noticeable is the choice of features while creating the data representation. In Table 1 we see the high number of finally selected features about attachments (in particular for spam data), route of emails by servers and other parts of header (e.g. senders). Such information was not available for Newsgroups, which may partly explain its lower classification results. For all data sets we observed an increase of classification accuracy when applying all these additional features. Using exclusively either the subject or the body features led to lower accuracy for all types of learned classifiers. For MarCas and NewsGroups data the subject features were the least useful while for For DWeiss data they were more important (perhaps it is connected with spam properties).

These observations are confirmed by the analysis of classifiers performance in separate classes – see Table 3. It is clearly illustrated by changes of the F-measure. Although these are results for 3-NN classifier, quite similar numbers were obtained for J4.8 decision trees, while slightly worse again for Naïve Bayes. Considering other measures it seems that additional features affected recall more than precision. For some folders in MarCas data the influence of using all features is really high, e.g. see recall in categories Humor, pJug, Work, Study or Family. On the other, it had lower influence on precision, e.g. in folder BooBoo, pJuG or Study. For NewsGroups, the use of the subject features only did not contribute too much to recall and precision, while the body features had also more influence on recall than on precision. Depending on the given folder, some results for the body features were quite comparable to the use of all features. For detecting

spams, i.e. DWeiss data set, the use of all features increased both precision and recall, and its influence was stronger for the spam category than regular emails. For precision the subject's features had higher influence on detecting spam while the body features were quite informative for classifying legitimate mails.

Comparing our results to related works on folder categorization of English emails, we could notice that researchers stress mainly the role of the body parts, see e.g. [3]. Although some others say that features coming from recipient or sender email parts are also useful (e.g. [4]), there is still lack of experimental evaluations. An exception is the recent study [3], the influence of different feature subsets on SVM classifiers was studied with conclusion that combining additional features improves the F measure and the body or sender features are more suitable than ones coming from the subject part.

Finally, we have to admit that we used only 3 data sets, so we should be cautious with making general conclusions. On the other hand, according to our best knowledge this research problem has not be examined yet in the context of Polish language and there have been no ready data sets. So, our original contribution is collecting at least these three data sets having different characteristics and making experiments with evaluating usefulness of various ways of constructing data representations. Saying this we risk conclusions that language processing techniques, as e.g. stemming, had smaller influence on the final classifiers than extending feature subsets by ones coming from the header part of an email and information about its attachments.

Acknowledgment: The research was supported from grant no. 3 T11C 050 26.

References

1. R. Bekkerman, A. McCallum, G. Huang: Automatic categorization of email into folders: Benchmark Experiments on Enron and SRI Corpora. U. Mass CIIR Report IR-418, 2004.
2. J. Clark, I. Koprinska, J. Poon: Linger – a smart personal assistant for e-mail classification. Proc. of the 13th Intern. Conf. on Artificial Neural Networks (ICANN'03), 2003, 274–277.
3. B. Klimt, Y. Yang: The enron corpus: a new dataset for email classification research. In Proc. of the ECML'04 Conference, 2004, 217–226.
4. G. Manco, E. Masciari, M. Ruffolo, A. Tagarelli: Towards an Adaptive Mail Classifier. Workshop "Apprendimento Automatico: Metodi ed Applicazioni", (AIIA 02). Siena, September 10-13, 2002
5. J. Stefanowski, D. Weiss: Carrot^2 and language properties in Web Search Results clustering. In Proc. of the 1st Atlantic Web Intelligence Conference AWIC-2003, LNCS 2663, 2003, 401-407.