The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Knowledge Extraction from Text: Technical Report WS-16-10

Effect of Part-of-Speech and Lemmatization Filtering in Email Classification for Automatic Reply

Rogerio Bonatti
Escola Politécnica
University of São Paulo
São Paulo, Brazil
rogerio.bonatti@usp.br

Arthur G. de Paula
Escola Politécnica
University of São Paulo
São Paulo, Brazil
arthur.paula@usp.br

Victor S. Lamarca
Escola Politécnica
University of São Paulo
São Paulo, Brazil
victor.lamarca@usp.br

Fabio G. Cozman
Escola Politécnica
University of São Paulo
São Paulo, Brazil
fgcozman@usp.br

Abstract

We study the automatic reply of email business messages in Brazilian Portuguese. We present a novel corpus containing messages from a real application, and baseline categorization experiments using Naive Bayes and Support Vector Machines. We then discuss the effect of lemmatization and the role of part-of-speech tagging filtering on precision and recall. Support Vector Machines classification coupled with non-lemmatized selection of verbs and nouns, adjectives and adverbs was the best approach, with 87.3% maximum accuracy. Straightforward lemmatization in Portuguese led to the lowest classification results in the group, with 85.3% and 81.7% precision in SVM and Naive Bayes respectively. Thus, while lemmatization reduced precision and recall, part-of-speech filtering improved overall results.

1 Introduction

Electronic mail is an ubiquitous mode of communication in personal and work life (Peng and Chan 2013; Chakrabarty 2014). Providing personalized answers to questions sent by email is not an easy task, particularly as the number of messages scales up (Richardson et al. 2001). Messages are written in natural language and may contain several questions concatenated in a single sentence, or even implicit questions, perhaps containing ambiguous terms. Automatic replies are particularly useful in enterprises and institutions that receive hundreds or thousands of emails per day regarding specific categories such as products or divisions.

Several techniques have been developed (Richardson et al. 2001; Buskirk Jr et al. 2003; Ayyadurai 2004) to automatically identify questions and intents in an email input, so as to either automatically answer questions or to forward the message to an expert. A common approach to text understanding is to classify incoming text into categories that are previously specified over the domain of interest. The first applications of email filtering appeared in the context of spam filtering, but classification methods can be applied to message filtering into user-defined folders, automatic forwarding to other addresses in companies with subject sectorization, and to automatic replies (Manning, Raghavan, and Shütze 2009). A major difference between spam detection and classification of email messages for automatic answering is the number of categories: while the former application has two categories, the latter application usually deals with dozens or even hundreds of potential classes depending on the complexity of the organization. Indeed this is the sort of challenge we face in our application.

In this paper we examine the problem of automatic email classification in multiple categories for business messages written in Brazilian Portuguese. Even though text understanding and binary email classification have been explored in the literature, very little work has been published on multi-categorical email classification for Portuguese. An exception is the work of Lima (2013), who describes work on binary email classification in Portuguese by exploring differences among multiple algorithms, but provides few details over the types of tests, datasets and results with classification over multiple categories.

We have been driven to this problem by observing the business automation needs concerning client service interaction in companies and institutions that receive hundreds of messages per day, in most part processed manually and inefficiently considering current natural language processing (NLP) technology. Our first goal was to build a corpus, containing business client interaction messages in Brazilian Portuguese, large enough for training / testing of statistical classification methods. Our second goal was to explore automatic email classification in Brazilian Portuguese with this dataset; first using Naive Bayes and Support Vector machines as a baseline for future study, and then by evaluating the impact of a lemmatizer pre-processing stage, and the impact of a part-of-speech tagger feature selector.

2 Background and Related Work

One can use machine learning algorithms to automatically learn classifiers based on training data that were previously classified by hand, in a supervised learning process (Manning, Raghavan, and Shütze 2009). Usually the accuracy of resulting classifiers is dependent upon the quantity of training data available (Thorsten 1999; Cohen, Carvalho, and Mitchell 2004). Often one combines labeled and unlabeled data (Li et al. 2014; Kiritchenko and Matwin 2001); in this paper we focus on supervised learning only.

Among the most accurate algorithms for text classification today are Support Vector Machines (SVM), Naive

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Bayes (NB) and k-Nearest-Neighbors (kNN), including hybrid approaches that can achieve greater precision than these methods separately (Pawar and Gawande 2012). SVM is one of the top performers for longer texts, but may present problems with shorter snippets (Wang and Manning 2012). SVM is usually implemented with linearly separable text in binary classification such as spam vs. ham, or sentiment analysis such as positive vs. negative. Multi-categorical applications are also common, and are usually solved by using a sequence of binary classifications of the type one-versus-rest (Liu and Yuan 2011).The NB method relies on the frequency of a word in the text. Because email data in the business context usually consists of relatively long sentences (large number of words in the vocabulary, i.e., high dimensional feature space), this paper focuses on the SVM and NB methods only, due to their robustness (Joachims 2001) to deal with such constraints.

Text classification algorithms typically take into account three types of features extracted from emails: unstructured text, categorical text and numerical data (Masciari, Ruffolo, and Tagarelli 2002). Unstructured text consists of the subject line and corpus, usually grouped in a ”bag of words”, while categorical data is well defined, and can be found in the sender and recipient domains for instance. Numerical data is related to message size and number of recipients. Additionally, feature selection filters may be applied to reduce noise in document classification and also to reduce the vocabulary used in computations.

Classifiers may use features based on word complexity, part-of-speech (POS) tags and presence of alphanumeric characters to enhance classification (Shams and Mercer 2013). A POS Tagger filter can be applied to the studied corpora to remove classes of words considered irrelevant or noise to text classification (such as verbs, nouns, adjectives, etc.). As discussed in literature (Salvetti, Lewis, and Reichenbach 2004), it may be advantageous to use POS tags in text classifiers, because, in many cases, information retrieval with POS tags improves the quality of the analysis (Losee 2001), and because it is a computationally inexpensive method to increase relevance in the training set.

Lemmatizers can also be used in text categorization to treat different variation of the same root words as one for statistical counting; for instance, to bring verbs to the infinitive form, and nouns to the singular and masculine form.

There are many examples of email and short text classification using machine learning algorithms in literature. For example, Klimt and Yang (2004) presented a email classification system in folders using the Enron Dataset with an F1 score near 70% using a SVM classifier. Chen et al. (2012) worked with microblog messages such as Twitter, classifying them into 6 categories like Sports, Business, etc., achieving both precision and recall close to 80%. Microblog messages are similar to emails in the sense that they use colloquial language and present relatively short sentences.

In Portuguese binary classification algorithms have achieved state-of-the-art levels. Silva (2009) and Moreira (2010) presented spam classifiers with true positive rates above 99%. Also in the field of short document analysis, Santos 2013) classified online product reviews as positive or negative with 78% precision and 81% recall.

Lima (2013) produced significant results on the topic of business email classification in Portuguese, comparing the performance of different classifiers on a set of emails labeled in folders. Lima presents F1 scores around 90% for binary classification in folders using kNN, and achieving 76% precision and 81% F1 for multi-topic classification with SVM. However, Lima provides few details on the reasons why SVM outperformed other classifiers in terms of the specific dataset characteristics that are not publicly available.

On the topic of feature selection for text classification, several papers are worth mentioning. In the micro blog context, Kouloumpis, Wilson, and Moore (2011) classified Twitter messages into positive or negative using multiple linguistic features such as separating words in n-grams, lexicon polarity and part-of-speech tags in different combinations. Results showed that using POS tags as a word feature decreased classification accuracy, going from about 65% F1 in the best case to approximately 55% F1 when POS tags are applied. Work by Batool et al. 2013) took a different approach of the use of filters: keywords were extracted from the text, and the best results were obtained with leaving only verbs and entities like hash tags in the text.

Pang, Lee, and Vaithyanathan (2002) tried another method in movie review sentiment analysis: comparing the performance of NB and SVM classifiers with datasets containing all parts-of-speech versus solely using adjectives for classification. Their results showed that despite the apparent expectation that adjectives contain most of the information relative to the positivity or negativity of a movie review, the vocabulary limitation actually decreased classification performance from 82% to 77% in accuracy.

3 Corpus collection

In this section we explain how we built our corpus.

To our knowledge, no public enterprise email corpus with multiple labeled categories is now available in Brazilian Portuguese, so it was necessary to partner with a company to run our experiments. We formed a partnership with Fundação Estudar, a non-profit organization in the field of education, that offers services such as student funding, prep courses and entrepreneurship workshops. They receive an average of 200 emails per day and agreed to share a database of 35,218 emails with us. All emails were written in Brazilian Portuguese, and were accumulated over a period of six months for both incoming and outgoing traffic. We chose not to collect data over a longer period, because the institution significantly changed its activities prior to a 6-month period, which could affect classification negatively due to changes in message categories.

Email is structured in tickets Messages in our corpus, as commonly seen in costumer relations services, are structured in *tickets*. A ticket corresponds to two or more email exchanges over the same topic. Typically, a ticket starts with a first email from a costumer asking a question, requesting technical support or sometimes giving information to the institution. Customer relations staff then reply the first email.

Nearly 75% percent of tickets, end in two interactions. Cases in which the same customer contacts the institution again through another email, an additional ticket is created to store the new conversation. We assembled 15,297 tickets in total.

Macros as classes for machine learning algorithm Fundação Estudar's customer relations staff, in their daily work of replying emails, use pre-written messages as the base for the responses to the most frequently received emails, over which they make small changes. The pre-written messages, called *macros*, represent 120 types of frequently received emails on the products or product's subcategories from Fundação.

Correctly classifying an incoming email into one of the classes in the subset is necessary to reply to it automatically. The actual answer employed by Fundação Estudar is a slightly modified version of the macro.

Email Labeling in Categories Tickets with more than three emails were not considered in our analysis. Discussions with staff responsible for answering emails gave us a better understanding of patterns in tickets, and we noticed that if a ticket had more than three emails, in most cases it was either the case that the response was not appropriate and the costumer needed another interaction, or that the topic was not clearly classifiable within the pre-determined categories, i.e., it could not be answered by a pre-written reply. Therefore we considered tickets containing two or three email messages: one question email, one response email and one optional thank-you email.

After the first triage, 11,410 tickets out of the original 15,297 remained. The next step of preparation of the corpus was the creation of our labeled data, obtained by labeling the remaining tickets within the classes, i.e., determining which macro could reply each of the emails. This was accomplished by comparing the institution's answers with the pre-defined responses in search for matches.

On average there were 28 emails per class after matching the macros to the actual replies. The distribution of emails per class is depicted in Figure 1. We separated 8 of the top 12 categories in number of emails for analysis, obtaining a total of at least 42 emails per class. 4 of these categories were discarded because they were generic answers that could refer to a variety of situations in the context of Fundação Estudar. 7 of the 8 chosen categories had a number of emails larger than 42, but to balance the classes vocabulary range and improve classification performance we selected as much emails as the eighth class.

Other Text and Email Corpora Other Portuguese language databases of manually annotated categories could be found, such as Linguateca (Freitas et al. 2010) and Floresta Sintática (Santos and Rocha 2004), but they do not contain email messages. The work of Lima (2013) contained an email corpus extracted from a private company in Portuguese, but it was not publicly available. Of course, in English there are several public corpora of labeled text belonging to more than 2 categories, such as the Reuters-21578 (Crawford, Kay, and McCreath 2001) corpus for news classification and the Enron (Klimt and Yang 2004) corpus for email classification, but the scope of our work was email classification in Portuguese.

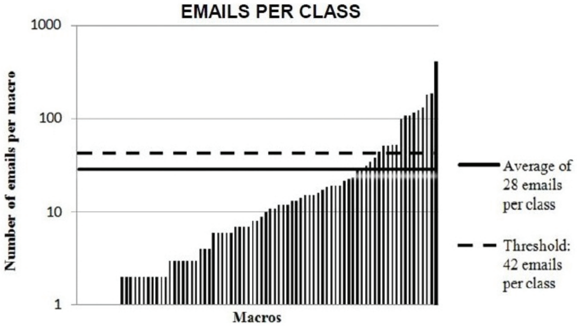

Figure 1: Distribution of emails per class .

4 Corpus Processing, and the Effect of Filtering

In this section we explain how we processed our email corpus to prepare the datasets used in the experiments. We also describe the techniques used to filter text (lemmatization and part-of-speech tagging), and the results with our corpus.

Text and Email Corpora in Literature We used different techniques to process the training corpus so as to assess the impact on recall and precision of removing certain parts-of-speech and of lemmatizing the text of the messages. The first dataset preparation was to use a Brazilian Portuguese lemmatizer (Coelho 2007) to bring verbs to infinitive form and nouns and adjectives to masculine and singular form. After this stage, the two corpora created, raw and lemmatized, were split into 18 groups by removing certain parts-of-speech and retaining others, with the use of a POS-Tagger for Brazilian Portuguese (Fonseca and Rosa 2013). Filter configurations are shown in Table 1.

Naive Bayes and SVM Classifiers In our experiments two different classifiers were utilized: Naive Bayes and Support Vector Machines. Distinct configurations for each of these algorithms were chosen taking into account the characteristics of our dataset.

For NB we opted for the multinomial configuration with Inverse Document Frequency (IDF) weighing for the vocabulary. These settings were chosen after literature review (Manning, Raghavan, and Shütze 2009; Metsis, Androutsopoulos, and Paliouras 2006) and preliminary tests with our dataset that showed its performance superior in relation to other options.

For the SVM classifier a linear kernel was used; the linear kernel's superior performance for text has been shown by Joachims (2001) and by Hsu, Chang, and Lin (2003). Preliminary experiments showed that using IDF weighing diminished performance with SVM, therefore IDF was not used in the experiments.

Evaluation of Lemmatization and Part-of-Speech Filtering on Classifier Performance Table 1 presents the ef-

Table 1: Effect of lemmatization and POS filtering on precision, recall and F1.

Datasets		**Naive Bayes**			**SVM**		
Lemmatizer	**POS-Tagger Filter**	**PR (%)**	**REC (%)**	**F1 (%)**	**PR (%)**	**REC (%)**	**F1 (%)**
No	No	82.7	81.8	81.7	85.6	85.1	85.3
No	Verbs and nouns without participle	84.1	83.3	83.0	86.9	86.0	86.3
No	Verbs and nouns only	84.9	83.9	83.7	87.1	86.6	86.8
No	Verbs, nouns and adjectives	85.4	84.5	84.4	86.1	85.7	85.9
No	**Verbs, nouns, adjectives and adverbs**	**84.7**	**84.2**	**83.9**	**87.5**	**87.2**	**87.3**
No	Verbs, nouns, and relative pronouns	84.3	83.3	83.1	87.1	86.6	86.8
No	Verbs, nouns and conjunctions	84.4	83.6	83.3	86.3	85.7	85.9
No	Verbs, nouns and adverbs	85.1	84.2	83.9	86.7	86.3	86.4
Yes	No	83.0	82.1	82.1	85.1	84.5	84.6
Yes	Verbs and nouns without participle	83.4	82.4	82.4	84.7	84.2	84.3
Yes	Verbs and nouns only	83.8	82.7	82.8	83.8	83.3	83.4
Yes	Verbs, nouns and adjectives	83.6	82.4	82.4	84.5	83.9	84.0
Yes	Verbs, nouns, adjectives and adverbs	84.3	83.0	83.1	86.0	85.4	85.5
Yes	Verbs, nouns, and relative pronouns	83.5	82.4	82.5	83.2	82.7	82.8
Yes	Verbs, nouns and conjunctions	83.7	82.7	82.8	84.9	84.5	84.5
Yes	Verbs, nouns and adverbs	82.5	81.5	81.5	85.0	84.5	84.6

fect of the POS-Tagger filter and of the lemmatizer in precision, recall and F1 measurements with our different training and test data. Comparing both classifiers among all filters, the highest precision achieved was 87.5%, recall 87.2% and F1 87.3%, for the training set containing verbs, nouns, adjectives and adverbs with unlemmatized emails and using linear-kernel SVM without IDF weighing. The results show that the lemmatizer reduces performance of the classifier, whereas the POS-Tagger improves it.

5 Conclusions

We successfully built a corpus of email messages in Brazilian Portuguese. That was accomplished in association with Fundação Estudar, a non-profit organization in education that provided us with their email logs.

Based on the corpus created, we produced a study of email classification. We implemented Naive Bayes and Suport Vector Machine email classifiers and tested precision, recall and F1 statistics for the use of a part-of-speech filter and for the use of a lemmatizer. The values of precision and recall obtained in our experiments are higher than what is seen in literature for email classification, or even general text classification. Our classifier reached precision, recall and F1 of 87.3%, above the range of 70 to 80% recall presented by Androutsopoulos et al. (2000) in binary classification for spam and ham. On multi-category classification, Dewdney, VanEss-Dykema, and MacMillan (2001) tested different algorithms for seven very distinct categories and obtained, approximately, recall of 76% and precision of 80%. Chen et al. (2012), who classified micro blog text within ten categories, reached 87% for both precision and recall.

We would like to emphasize the following conclusions:

1. Colloquial speech affected performance negatively. One characteristic of our corpus that reduces performance is the informality in email messaging. For example, when compared to a collection of newspaper articles as Reuters-21578 (Crawford, Kay, and McCreath 2001) that has much more vocabulary per text, longer texts and more formal language use, our corpus presents greater challenges for classification as these characteristics have great effect on the machine learning algorithm. In informal language, the reduced variety of words that are used results in a higher chance of finding two emails that have the same words and belong to different classes.

2. The use of a lemmatizer was not beneficial.

To explain this experimental result, consider that lemmatization removes information that the words' inflections carry, such as verb tenses. An analogy can be made with a three-dimensional castle of cards. Lemmatizing a word would be the same as taking a photograph of the castle from the top: from the photo, you can still see some cards, but you no longer understand they form a castle, nor see the rest if them. Lemmatizing the words is losing a dimension of it, just like in the castle of cards. In our case as well as in our analogy, the dimension we lose causes loss in explanatory power.

3. Part-of-speech filtering did improve classification. The experiments we carried out showed significant increase in performance of the classifier for POS-filtered datasets, which suggests that, in our context, nouns and verbs are the most significant parts-of-speech for the classification.

A possible explanation for this phenomenon comes from the retained POS having better defined patterns for each class, considering our dataset size. The parts-of-speech removed (prepositions, conjunctions, pronouns, etc.) would then, act as noise in the classification.

Removing certain POS is reducing the information carried by the models for classification as well as using the lemmatizer, but for the POS, the filtered parts did not add relevant information to the classification. This observation is specific for our dataset in terms of both context and size.

In the context of email classification for costumer relations, nouns and verbs appear to carry the most relevant information, which may not be true for text classification in other contexts. In sentiment analysis, for example, adjectives and adverbs are likely to have greater importance.

6 Acknowledgments

We thank Fundação Estudar for providing all the data we used.

References

Androutsopoulos, I.; Paliouras, G.; Karkaletsis, V.; Sakkis, G.; Spyropoulos, C. D.; and Stamatopoulos, P. 2000. Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach. *Proceedings of the workshop Machine Learning and Textual Information Access* (September 2000):1–12.

Ayyadurai, S. 2004. System and Method for Content- Sensitive Automatic Reply Message Generation for Text-Based Asynchronous Communications.

Batool, R.; Khattak, A. M.; Maqbool, J.; and Lee, S. 2013. Precise tweet classification and sentiment analysis. *2013 IEEE/ACIS 12th International Conference on Computer and Information Science, ICIS 2013 - Proceedings* 461–466.

Buskirk Jr, M. C.; Damerau, F. J.; Johnson, D. H.; and Raaen, M. 2003. Machine Learning Based Electronic Messagind System.

Chakrabarty, A. 2014. An Optimized k-NN Classifier based on Minimum Spanning Tree for Email Filtering. *Business and Information Management (ICBIM)* 47–52.

Chen, Y.; Li, Z.; Nie, L.; Hu, X.; Wang, X.; Chua, T.-s.; and Zhang, X. 2012. A Semi-Supervised Bayesian Network Model for Microblog Topic Classification. *Coling* 1(December):561–576.

Coelho, A. R. 2007. Stemming para a língua portuguesa: estudo, análise e melhoria do algoritmo RSLP. Master's thesis, Universidade Federal do Rio Grande do Sul.

Cohen, W. W.; Carvalho, V. R.; and Mitchell, T. M. 2004. *Learning to Classify Email into "Speech Acts"*, volume 4.

Crawford, E.; Kay, J.; and McCreath, E. 2001. Automatic Induction of Rules for e-mail Classification. In *Sixth Australasian Document Computing Symposium*.

Dewdney, N.; VanEss-Dykema, C.; and MacMillan, R. 2001. The Form is the Substance: Classification of Genres in Text. In *Proceedings of the workshop on Human Language Technology and Knowledge Management*, 1–8. Association for Computational Linguistics.

Fonseca, E. R., and Rosa, G. 2013. Mac-Morpho Revisited: Towards Robust Part-of-Speech Tagging. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 98–107.

Freitas, C.; Mota, C.; Santos, D.; Oliveira, H. G.; and Carvalho, P. 2010. Second HAREM : Advancing the State of the Art of Named Entity Recognition in Portuguese. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (3):3630–3637.

Hsu, C.-W.; Chang, C.-C.; and Lin, C.-J. 2003. A Practical Guide to Support Vector Classification. *BJU international* 101(1):1–16.

Joachims, T. 2001. *Learning to Classify Text Using Support Vector Machines*, volume 29.

Kiritchenko, S., and Matwin, S. 2001. Email Classification with Co-Training. *Proceedings of the 2001 Conference of the Centre for Advanced Studies on Collaborative research* 8.

Klimt, B., and Yang, Y. 2004. The Enron corpus: A new dataset for Email Classification Research. *Machine Learning: ECML 2004* 217–226.

Kouloumpis, E.; Wilson, T.; and Moore, J. 2011. Twitter Sentiment Analysis : The Good the Bad and the OMG ! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* 538–541.

Li, W.; Meng, W.; Tan, Z.; and Xiang, Y. 2014. Towards Designing an Email Classification System Using Multi-view Based Semi-Supervised Learning. *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications* 174–181.

Lima, M. J. A. 2013. Classificação Automática de Emails. Master's thesis, Universidade do Porto.

Liu, Y., and Yuan, M. 2011. Reinforced Multicategory Support Vector Machines. *Journal of Computational and Graphical Statistics* 20(4):901–919.

Losee, R. M. 2001. Natural language processing in support of decision-making: Phrases and part-of-speech tagging. *Information Processing and Management* 37(6):769–787.

Manning, C. D.; Raghavan, P.; and Shütze, H. 2009. *An Introduction to Information Retrieval*. Cambridge: Cambridge UP, c edition.

Masciari, E.; Ruffolo, M.; and Tagarelli, A. 2002. Towards an Adaptive Mail Classifier. *Italian Association for Artificial Intelligence Workshop Su Apprendimento Automatico: Metodi ed Applicazioni* (August).

Metsis, V.; Androutsopoulos, I.; and Paliouras, G. 2006. Spam filtering with Naive Bayes -Which naive bayes? *Ceas* 9.

Moreira, M. S. 2010. Detecção De Mensagens Não Solicitadas Utilizando Mineração De Textos. Master's thesis, COPPE UFRJ.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* 79–86.

Pawar, P. Y., and Gawande, S. H. 2012. A Comparative Study on Different Types of Approaches to Text Categorization. *International Journal of Machine Learning and Computing* 2(4):423–426.

Peng, J., and Chan, P. P. K. 2013. Revised Naive Bayes Classifier for Combating the Focus Attack in Spam Filtering. In *2013 International Conference on Machine Learning and Cybernetics*, 14–17.

Richardson, K. D.; Greif, J.; Buedel, D.; and Aleksandrovsky, B. 2001. System And Method For Message Process And Response.

Salvetti, F.; Lewis, S.; and Reichenbach, C. 2004. Automatic Opinion Polarity Classification of Movie. *Colorado Research in Linguistics* 17(1):2.

Santos, D., and Rocha, P. 2004. CHAVE : topics and questions on the Portuguese participation in CLEF. *Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop* 639—-648.

Santos, F. 2013. Mineração de opinião em textos opinativos utilizando algoritmos de classicação. Master's thesis, Universidade de Brasilia.

Shams, R., and Mercer, R. E. 2013. Classifying spam emails using text and readability features. *Proceedings - IEEE International Conference on Data Mining, ICDM* 657–666.

Silva, A. M. D. 2009. Utilização De Redes Neurais Artificiais Para Classificação De Spam. Masters, Centro Federal De Educação Tecnológica De Minas Gerais.

Thorsten, J. 1999. Transductive Inference for Text Classification Using Support Vector Machines. *ICML* 99:200–209.

Wang, S., and Manning, C. D. C. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. *ACL 12 Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2* 94305(1):90–94.