Feature Selection for Short Text Classification using Wavelet Packet Transform

Anuj Mahajan
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi 110016
India
anujmahajan.iitd@gmail.com

Sharmistha, Shourya Roy
Xerox Research Centre India
Bangalore - 560103
India
{*sharmistha.jat,shourya.roy*}*@xerox.com*

Abstract

Text classification tasks suffer from curse of dimensionality due to large feature space. Short text data further exacerbates the problem due to their sparse and noisy nature. Feature selection thus becomes an important step in improving the classification performance. In this paper, we propose a novel feature selection method using Wavelet Packet Transform. Wavelet Packet Transform (WPT) has been used widely in various fields due to its efficiency in encoding transient signals. We demonstrate how short text classification task can be benefited by feature selection using WPT due to their sparse nature. Our technique chooses the most discriminating features by computing inter-class distances in the transformed space. We experimented extensively with several short text datasets. Compared to well known techniques our approach reduces the feature space size and improves the overall classification performance significantly in all the datasets.

1 Introduction

Text classification task consists of assigning a document to one or more classes. This can be done using machine learning techniques by training a model with labelled documents. Documents are usually represented as vectors with a variety of techniques like bag-of-words(unigram, bigram), TFIDF representation, etc. Typically, text corpora have very high dimensional document representation equal to the size of vocabulary. This leads to curse of dimensionality[1] in machine learning models, thereby degrading the performance.

Short text corpora, like SMS, tweets, etc., in particular suffer from sparse high dimensional feature space, due to large vocabulary and short document length. To give an idea as to how these factors affect the size of the feature space we compare Reuters with Twitter data corpus. In Reuters-21578 corpus there are approximately 2.5 Million words in total and 14506 unique vocabulary entries after standard preprocessing steps (which is the dimensionality of the feature space). However, Twitter 1 corpus, we used for experiments has approximately 15,000 words in total and feature space size of 7423 words. Additionally, the average length of an English tweet is 10-12 words whereas the average length of a document in Reuters-21578 news classification corpus is 200 words. Therefore, the dimensionality is extremely high even for small corpora with short texts. In addition, the average number of words in a document is significantly less in short text data leading to higher sparsity of feature space representation of documents.

Owing to this high dimensionality problem, one of the important steps in text classification workflows is feature selection. Feature selection techniques for traditional documents have been aplenty and a few seminal survey articles have been written on this topic (Blitzer, 2008). In contrast, for short text there is much less work on statistical feature selection but more focus has gone to feature engineering towards word normalization, canonicalization etc. (Han and Baldwin, 2011).

In this paper, we propose a dimensionality reduction technique for short text using Wavelet packet transform called Improvised Adaptive Discriminant Wavelet Packet Transform (IADWPT). IAWDPT does dimensionality reduction by selecting discriminative features (wavelet coefficients) from the Wavelet Packet Transform (WPT) representation. Short text data resembles transient signals in vector representation and WPT encodes transient signals (signals lasting for very short duration) well (Learned and Willsky, 1995), using very few coefficients. This leads to considerable decrease in the dimensionality of the feature space along with increase in classification accuracy. Additionally, we optimise the procedure to select the most discriminative features from WPT representation. To the best of our knowledge this is the first attempt to apply an algorithm based on wavelet packet transform to the feature selection in short text classification.

2 Related Work

Feature selection has been widely adopted for dimensionality reduction of text datasets in the past. Yiming Yang et al. (Yang and Pedersen, 1997) performed a comparative study of some of these methods including, document frequency, information gain(IG), mutual

[1] https://en.wikipedia.org/wiki/Curse_of_dimensionality

Proceedings of the 19th Conference on Computational Language Learning, pages 321–326,
Beijing, China, July 30-31, 2015. ©2015 Association for Computational Linguistics

information(MI), χ^2-test(CHI) and term strength(TS). They concluded that IG and CHI are the most effective in aggressive dimensionality reduction. The mRMR technique proposed by (Peng et al., 2005) selects the best feature subset by increasing relevancy of feature with target class and reducing redundancy between chosen features.

Wavelet transform provides time-frequency representation of a given signal. The time-frequency representation is useful for describing signals with time varying frequency content. Detailed explanation of wavelet transform theory is beyond the scope of this paper. For detailed theory, refer to Daubechies (Daubechies, 2006; Daubechies, 1992; Coifman and Wickerhauser, 2006) and Robi Polikar (Polikar,). First use of wavelet transform for compression was proposed by Ronald R Coifman et al. (Coifman et al., 1994). Hammad Qureshi et al. (Qureshi et al., 2008) proposed an adaptive discriminant wavelet packet transform(ADWPT) for feature reduction.

In past wavelet transform has been applied to natural language processing tasks. A survey on wavelet applications in data mining (Li et al., 2002), discusses the basics and properties of wavelets which make it a very effective technique in Data Mining. CC Aggarwal (Aggarwal, 2002) uses wavelets for strings classification. He notes that wavelet technique creates a hierarchical decomposition of the data which can capture trends at varying levels of granularity and thus helps classification task with the new representation. Geraldo Xexeo et al. (Xexeo et al., 2008) used wavelet transform to represent documents for classification.

3 Wavelet Packet Transform for Short-text Dimensionality Reduction

Feature selection performs compression of feature space to preserve maximum discriminative power of features for classification. We use this analogy to do compression of document feature space using Wavelet Packet Transform. Vector format(e.g. dictionary encoded vector) representation of a document is equivalent to a digital representation. This vector format can then be processed using wavelet transform to get a compressed representation of the document in terms of wavelet coefficients. Document features are transformed into wavelet coefficients. Wavelet coefficients are ranked and selected based on their discrimination power between classes. Classification model is trained on these highly informative coefficients. Results show a considerable improvement in model accuracy using our dimensionality reduction technique.

Typically, vector representation of short text will have very few non-zero entries due to short length of the documents. If we plot count of each word in the dictionary on y-axis v/s distinct words on x-axis. Just like transient signals, the resulting graph will have very few spikes. Transient signals last for a very little time in the whole duration of the observation. (Learned and Willsky, 1995) show the efficacy of wavelet packet transform in representing transient signal. This motivates our use of Wavelet Packet Transform to encode short text.

Wavelet transform is a popular choice for feature representation in image processing. Our approach is inspired by a related work by (Qureshi et al., 2008). They propose Adaptive Discriminant Wavelet Packet Transform (ADWPT) based representation for meningioma subtype classfication. ADWPT obtains a wavelet based representation by optimising the discrimination power of the various features. Proposed technique IADWPT differs from ADWPT in the way discriminative feature are selected. Next section provides details about the proposed approach IADWPT.

4 IADWPT - Improvised Adaptive Discriminant Wavelet Packet Transform

This section presents the proposed short text feature selection technique IADWPT. IADWPT uses wavelet packet transform of the data to extract useful discriminative features from the sub-bands at various depths.

Natural language processing tasks usually represent their documents in dictionary encoded bag-of-words representation. This numerical vector representation of a document is equivalent to signal representation. In order to get IADWPT representation of the document following steps should be computed:

1) Compute full wavelet packet transform of the document vector representation.

2) Compute the discrimination power of each coefficient in wavelet packet transform representation.

3) Select the most discriminative coefficients to represent all the documents in the corpus.

Once the 1-D wavelet transform is computed at a desired level l, wavelet packet transform (WPT) produces 2^l different sets of coefficients (nodes in WPT tree). These coefficients represent the magnitude of various frequencies present in the signal at a given time. We select the most discriminative coefficients to represent all the documents in the corpus by calculating the discriminative power of each coefficient.

The classification task consists of c classes with d documents. 1-D Wavelet Packet Transform of the d_k^{th} document yields l levels with f sub bands consisting of m coefficients in each sub band. $x_{m,f,l}$ represent the coefficients of Wavelet Packet Transform. Following terms are defined for Algorithm 1.

- probability density estimates ($S_{m,f,l}$) of a particular sub-band in a level l a training sample document d_k^i of a given Class c_i is given by:

  $S_{m,f,l}^k = \frac{(x_{m,f,l}^k)^2}{\sum_j (x_{j,f,l}^k)^2}$

  Here, $x_{m,f,l}$ is the m^{th} coefficient in f^{th} subband of l^{th} level of document d_k. Where, j varies

Algorithm 1 IADWPT Algorithm for best discriminative feature selection

```
for all classes $C$ do
    Calculate Wavelet Packet Transform for all the documents $d_k$ in class $c_i$
    for all Documents $d_k$ do
        Calculate probability density estimates $S^k_{m,f,l}$
    end for
    for all Levels of WPT $l$ and their sub bands $f$ do
        for all Wavelet Packet Transform Coefficients $m$ in subband $f$ do
            Calculate average probability density $A^{c_i}_{m,f,l}$
        end for
    end for
end for
for all Class Pairs $c_a, c_b$ do
    Calculate discriminative power $D^{a,b}_{m,f,l}$
end for
Select top $m'$ coefficients for representing documents in corpus
```

over the length of sub-band. Wavelet supports vary with the bands, Normalization ensures that the feature selection done gives uniform weightage to the features from different bands. This step calculates the normalised value of coefficients in a sub-band.

- Average probability density ($A^{c_i}_{m,f,l}$) estimates are derived using all the training samples $d_k$in a given class c_i.

  $A^{c_i}_{m,f,l} = \frac{\sum_k S^k_{m,f,l}}{d}$

  for, coefficient m in sub-band f for class c_i and d is the total number of documents in the class. k varies over the number of documents in the class. It measures the average value of a coefficient in all the documents belonging to a class.

- Discriminative power ($D^{a,b}_{m,f,l}$) of each coefficient in l^{th} level's f^{th} sub band's m^{th} coefficient, between classes a and b is defined as follows:

  $D^{a,b}_{m,f,l} = |\sqrt{A^a_{m,f,l}} - \sqrt{A^b_{m,f,l}}|$

  Discriminative power is the hellinger distance between the average probability density estimates of a coefficient for the two classes. It quantifies the difference in the average value of a coefficient between a pair of classes. More the difference, better the discriminative power of the coefficient. Thus discriminative features tend to have a higher average probability density in one of the classes whereas redundant features cancel out in taking the difference in computing the distance. (Rajpoot, 2003) have shown efficacy of Hellinger distance applied to ADWPT.

Selecting coefficients with greater discriminative power helps the classifier perform well. Full algorithm is mentioned in algorithm 1.

Multi class classification can then be handled in this framework using one-vs-one classification. We select the top m' features from the wavelet representation for representing the data in the classification task. Time complexity of the algorithm is polynomial. The method is based on adaptive discriminant wavelet packet transform (ADWPT) (Qureshi et al., 2008). Therefore, we name it as improvised adaptive discriminant wavelet packet transform (IADWPT). ADWPT uses best basis for classification which is a union of the various subbands selected that can span the transformed space, so noise is still retained in the signal whereas IADWPT selects coefficients from the sub-band having maximal discriminative power thus improving the classification results. As opposed to ADWPT, IADWPT is a one way transform, original signal cannot be recovered from the transform domain. Experimental results confirm that IADWPT performs better than ADWPT in short text datasets.

4.1 IADWPT Example

Figure 1 Gives intuition of the workings of IADWPT transform. Uppermost graph displays the Average probability density in positive class samples. Middle graph in the Figure 1 shows the Average probability density in negative class samples. These two energy values are then subtracted, resulting values are shown in the bottom most component of Figure 1. Peak positive and negative values in the bottommost graph represent the most discriminant features. Absolute value of the discriminative power can then be used to select the most discriminative features to represent each document in corpus.

5 Experiments and Results

We used multiple short text datasets to prove efficacy of proposed algorithm against state of the art algorithms for feature selection.

1) **Twitter_1**: This dataset is a part of the SemEval 2013 task B dataset (Nakov et al.,) for two class sentiment classification. We gathered 624 examples in positive and negative class each for our experiments.

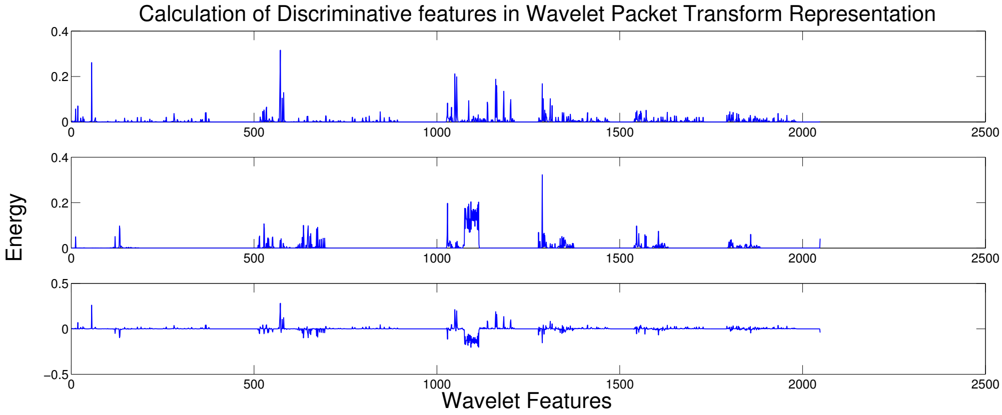

Figure 1: x-axis represents the wavelet packet transform coefficients, y-axis represents amplitude. From top to bottom, 1) Average probability density $A^{a}_{m,f,l}$value of coefficients in positive class a), 2) Average probability density $A^{b}_{m,f,l}$value of coefficients in negative class (b), 3) Difference between $A^{a}_{m,f,l}$ and $A^{b}_{m,f,l}$

Table 1: CLASSIFICATION RESULTS - SUPPORT VECTOR MACHINE (SVM) AND LOGISTIC REGRESSION (LR)

Dataset	Baseline	MI-avg	χ^2	PCA	ADWPT	IADWPT
Classification accuracy - SVM						
Twitter 1	47.04	47.57	45.62	59.28	46	**63.44**
SMS Spam 1 - HAM Accuracy	99.82	99.71	99.77	99.87	99.63	**99.94**
SMS Spam 1 - SPAM Accuracy	83.10	83.32	82.96	83.81	83.57	**83.92**
SMS Spam 2 - HAM Accuracy	55.2	56.31	55.62	81.12	54.1	**87.7**
SMS Spam 2 - SPAM Accuracy	46.6	46.7	46.49	92.39	47.3	**99.42**
Total dimensions in best classification accuracy result - SVM						
Twitter 1	7423	2065	515	540	7423	23
SMS Spam 1	9394	3540	550	750	9394	815
SMS Spam 2	10681	2985	490	855	1068	250
Classification accuracy - Logistic Regression						
Twitter 1	75.8	74.97	75.21	76.28	68.2	**76.72**
SMS Spam 1 - HAM Accuracy	97.91	94.67	95.28	98.71	98.03	**99.61**
SMS Spam 1 - SPAM Accuracy	**95.48**	85.34	86.37	91.37	82.2	87.54
SMS Spam 2 - HAM Accuracy	96.02	89.54	92.76	71.21	95.09	**98.5**
SMS Spam 2 - SPAM Accuracy	91.2	88.37	91.38	89.15	92.2	**94.51**
Total dimensions in best classification accuracy result - Logistic Regression						
Twitter 1	7423	5600	3250	3575	7423	2749
SMS Spam 1	9394	7545	1755	2350	9394	1680
SMS Spam 2	10681	6550	3000	3050	10681	9981

2) **SMS_Spam_1**: UCI spam dataset (Almeida,) consists of 5,574 instances of SMS classified into SPAM and HAM classes. SPAM class is defined as messages which are not useful and HAM is the class of useful messages. We compare our results with the results they published in their paper (Almeida et al., 2013). Therefore, we followed the same experiment procedure as cited in the paper. First 30% samples were used in train and the rest in test set as reported in the paper.

3) **SMS_Spam_2**: The dataset was published by Yadav et al. (Yadav et al., 2011). It consists of 2000 SMS, 1000 SPAM and 1000 HAM messages. Experiment settings are same as that of dataset SMS_Spam_1.

The goal of our experiments is to examine the effectiveness of the proposed algorithm in feature selection for short text datasets. We measure the effectiveness of the feature selection technique with respect to the increase in accuracy in the final machine learning task. Our method does not depend on a specific classifier used in the final classification. Therefore, we used

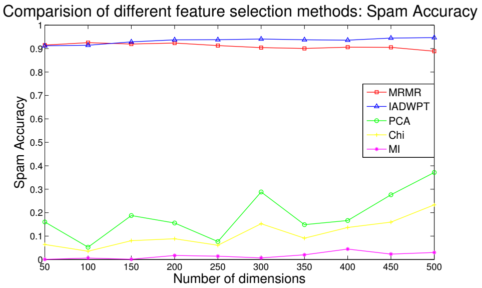

Figure 2: Spam Classification Accuracy comparison across various feature selection algorithm for SMS Spam 2 dataset

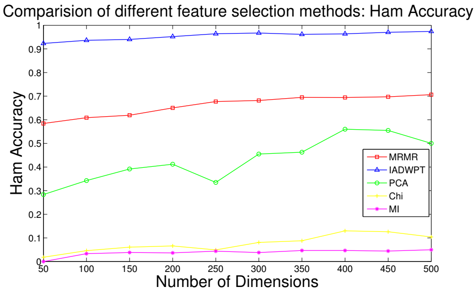

Figure 3: Ham Classification Accuracy comparison across various feature selection algorithm for SMS Spam 2 dataset

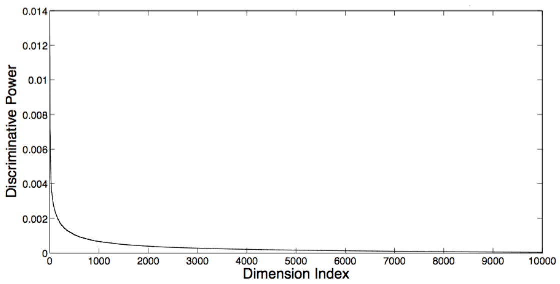

Figure 4: Plot of Discriminative Power ($D^{a,b}_{m,f,l}$) arranged in descending order

popular classifiers like Support Vector Machine (RBF kernel with grid search) (Cortes and Vapnik, 1995) and Logistic Regression with and without dimensionality reduction for unigram representation to benchmark the performance. All the experiments were done with 10 fold cross validation and grid search on C parameter. We report results with respect to classification accuracy which is measured as $\frac{\#correctly\ classified\ datapoints}{\#total\ datapoints}$.

We conducted detailed experiments comparing IADWPT using Coiflets of order 2 with other feature selection techniques such as PCA, Mutual Information, χ^2, mRMR (Peng et al., 2005) and ADWPT (Qureshi et al., 2008). Results are reported in Table 1. The table reports best accuracy values and respective feature set size selected by the technique. It can be observed that IADWPT gives best accuracy in most of the cases with very few features.

We compared performance of our algorithm with mRMR. Results for SMS_Spam_2 dataset are shown in Figure 2 and Figure 3. The plots prove efficacy of our algorithm versus state of the art mRMR algorithm. mRMR technique could not finish execution for the rest of the datasets. It can also be observed from results in Table 1 and Figure 1,2 that performance of feature selection algorithms follow consistent pattern in short text. Following is observed order of performance of algorithms in decreasing order, IADWPT, mRMR, PCA, Chi Square, MI. Further, it is observed that IADWPT performs well at feature selection without losing discriminative information, even when the dimensionality of feature space is reduced to as far as 1/40th of original feature space and steadily maintains the accuracy as dimensionality is reduced, which makes it a suitable technique for aggressive dimensionality reduction. This also helps in learning ML (machine learning) models faster due to reduced dimensionality. We plotted the discrimination power of coefficients in each dataset. Plot suggested that very few coefficients contained most of the discriminative power. And, therefore just working with these coefficients can help in getting good accuracies resulting in aggressive dimensionality reduction. Results establish the effectiveness of IADWPT for applicability in compressing short text feature representation and reducing noise to improve classification accuracy.

5.1 IADWPT Effectiveness

Short text data is noisy and consists of many features which are irrelevant to the task of classification of data. IADWPT effectively gets rid of the noise in approximations(as Signal strength is greater than the noise), the feature selection step at the sub-band level as described in Algorithm 1, it enforces selection of good discriminative features and thus improves classifier accuracy, reducing feature space dimensionality at the same time. Features from sub-bands of the signal are chosen based on their discriminative power, therefore, the original signal information is lost and the transform is not reversible.

IADWPT gives good compression of data and without losing discriminative information, even when the dimensionality of space is reduced to as far as 1/40th of original feature space and is thus steadily maintaining the accuracy as dimensionality is reduced, which makes it a suitable technique for dimensionality reduction. This also helps learning machine learning models faster due to reduced dimensionality. Figure 4 shows the plot of Discriminative Power $D^{a,b}_{m,f,l}$ values for coefficients arranged in descending order for SMS_Spam_2 dataset. Other datasets displayed similar graph for Discriminative Power. From the figure it can be observed that few coefficients hold most of the discriminative power, and thus aggressive dimensionality reduction is possible with IADWPT algorithm. Results establish the effectiveness of IADWPT for applicability in compressing short text feature representation and reducing noise.

6 Conclusion and Future Work

In this paper, we have proposed IADWPT algorithm for effective dimensionality reduction for short text corpus. The algorithm can be used in a number of scenarios where high dimensionality and sparsity pose challenge. Experiments prove efficacy of IADWPT based dimensionality reduction for short text data. This technique can prove useful to a number of social media data analysis applications. In future, we would like to explore theoretical bounds on best number of dimensions to choose from wavelet representation.

References

Charu C. Aggarwal. 2002. On effective classification of strings with wavelets.

Tiago A Almeida. Sms spam collection v.1. `http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/`,[Online; accessed 10-September-2014].

T. Almeida, J. M. G. Hidalgo, and T. P. Silva. 2013. Towards sms spam filtering: Results under a new dataset. *International Journal of Information Security Science*.

John Blitzer. 2008. A survey of dimensionality reduction techniques for natural language. `http://john.blitzer.com/papers/wpe2.pdf`,[Online; accessed 10-September-2014].

R. R. Coifman and M. V. Wickerhauser. 2006. Entropy-based algorithms for best basis selection. *IEEE Trans. Inf. Theor.*, 38(2):713–718, September.

RonaldR. Coifman, Yves Meyer, Steven Quake, and M.Victor Wickerhauser. 1994. Signal processing and compression with wavelet packets. In J.S. Byrnes, JenniferL. Byrnes, KathrynA. Hargreaves, and Karl Berry, editors, *Wavelets and Their Applications*, volume 442 of *NATO ASI Series*, pages 363–379. Springer Netherlands.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September.

Ingrid Daubechies. 1992. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

I. Daubechies. 2006. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Inf. Theor.*, 36(5):961–1005, September.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rachel E. Learned and Alan S. Willsky. 1995. A wavelet packet approach to transient signal classification. *Applied and Computational Harmonic Analysis*, 2(3):265 – 278.

Tao Li, Qi Li, Shenghuo Zhu, and Mitsunori Ogihara. 2002. A survey on wavelet applications in data mining. *SIGKDD Explor. Newsl.*, 4(2):49–68, December.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter.

Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238.

Robi Polikar. Wavelet Tutorial. `http://users.rowan.edu/~polikar/wavelets/wttutorial.html`,[Online; accessed 10-January-2015].

Hammad Qureshi, Olcay Sertel, Nasir Rajpoot, Roland Wilson, and Metin Gurcan. 2008. Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification. In Dimitris N. Metaxas, Leon Axel, Gabor Fichtinger, and Gbor Szkely, editors, *MICCAI (2)*, volume 5242 of *Lecture Notes in Computer Science*, pages 196–204. Springer.

NM Rajpoot. 2003. Local discriminant wavelet packet basis for texture classification. In *SPIE Wavelets X*, pages 774–783. SPIE.

Geraldo Xexeo, Jano de Souza, Patricia F. Castro, and Wallace A. Pinheiro. 2008. Using wavelets to classify documents. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:272–278.

Kuldeep Yadav, Ponnurangam Kumaraguru, Atul Goyal, Ashish Gupta, and Vinayak Naik. 2011. Smsassassin: Crowdsourcing driven mobile-based system for sms spam filtering. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, HotMobile '11, pages 1–6, New York, NY, USA. ACM.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.