Using Compression to Identify Classes of Inauthentic Texts

Mehmet M. Dalkilic, Wyatt T. Clark, James C. Costello, Predrag Radivojac*
{dalkilic, wtclark, jccostel, predrag}@indiana.edu
School of Informatics, Indiana University, Bloomington, IN 47408

Abstract

Recent events have made it clear that some kinds of technical texts, generated by machine and essentially meaningless, can be confused with authentic, technical texts written by humans. We identify this as a potential problem, since no existing systems for, say the web, can or do discriminate on this basis. We believe that there are subtle, short- and long-range word or even string repetitions extant in human texts, but not in many classes of computer generated texts, that can be used to discriminate based on meaning. In this paper we employ universal lossless source coding to generate features in a high-dimensional space and then apply support vector machines to discriminate between the classes of authentic and inauthentic expository texts. Compression profiles for the two kinds of text are distinct—the authentic texts being bounded by various classes of more compressible or less compressible texts that are computer generated. This in turn led to the high prediction accuracy of our models which support a conjecture that there exists a relationship between meaning and compressibility. Our results show that the learning algorithm based upon the compression profile outperformed standard term-frequency text categorization on several non-trivial classes of inauthentic texts. Availability: `http://www.informatics.indiana.edu/predrag/fsi.htm`.

1 Introduction

When operating over a corpus of text there is a natural presumption that the text is meaningful. This presumption is so strong that neither the tools, like webpage search engines, nor the people who use them take into account whether, for example, a webpage conveys any meaning at all, even though the number of indexable webpages available is so large and growing [4]. And yet, a web search for the nonsensical sentence, "Colorless green ideas sleep furiously," yields scores of thousands of hits on Google, Yahoo, and MSN. Of course this is no ordinary sentence—it is Noam Chomsky's famous sentence that he constructed to illustrate that grammar alone cannot ensure meaning [10]. While the sentence is syntactically correct and can be parsed, it does not possess any real meaning. But the important point is that the sentence *is* meaningless and has become part of the searchable text indistinguishable from any other sentence.

Single sentences can seldom convey enough meaning and are therefore combined into texts or documents to provide some larger, more complex information. According to linguists, texts exhibit not only sentential structure, but also higher levels of structure, for example, the so-called *expository structure* that are meant to be *informative*, that is, scholarly, encyclopedic, and factual as opposed to, say, those intended for entertainment. These higher level distinctions can be somewhat problematic if taken too literally, but are useful nonetheless. We can take other perspectives too: there are global patterns that are *only* manifested when the text is examined in its entirety. For example, one kind of global text pattern is the adherence to a *topic*. Another example is *discourse*—the different kinds of meaning derived solely from the arrangement of sentences.

To make clear the class of problem we are interested in examining, we provide the following definitions:

Definition 1.1. *An* authentic *text (or document) is a collection of several hundreds (or thousands) of syntactically correct sentences such that the text* as a whole *is meaningful. A set of authentic texts will be denoted by $\mathcal{A}$ with possible sub- or superscripts.*

Definition 1.2. *An* inauthentic *text (or document) is a collection of several hundreds (or thousands) of syntactically correct sentences such that the text* as a whole *is not meaningful. A set of inauthentic texts will be denoted by $\mathcal{I}$ with possible sub- or superscripts.*

Now consider a scenario in which inauthentic texts are not human generated, but are automated and embellished further with figures, citations, and bibliographies. Without dedicated human scrutiny such texts can escape identification and easily become part of searchable texts of cyberspace. Such a scenario recently played out when an automated inauthentic text was accepted to a conference without formal review, although we are not aware of the mechanism that led

*To whom correspondence should be addressed.

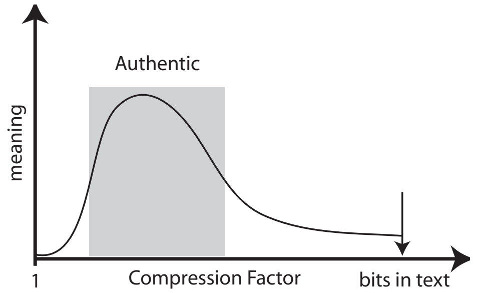

Figure 1: A stylized rendering of the relationship between compression and meaning in an expository text. Observe that the compression factor, the fraction of original document size to compressed size is between approximately one and the number of bits in the document.

to its acceptance. There are, in fact, scores of systems that generate inauthentic data and short texts, ranging from names, to email, and as demonstrated above, to text. While no direct numbers exist, we believe that currently most of the text in cyberspace is authentic. There is no reason to believe, however, that this will hold true in the future. Indeed, it is not too hard to foresee schemes in which "information pollution" (certainly one perspective of inauthentic texts) is produced to confound search results, perhaps obscuring authentic texts to the point of uselessness–Google spamming is one such kind of scheme.

We can identify what we believe to be an interesting problem: *create a classifier that distinguishes between authentic and computer generated inauthentic texts given a set containing both.* Clearly, we cannot make direct use of syntax as eluded to above. Furthermore, while text mining is gaining a lot of attention, no formal components of any of these models seek to either identify or even distinguish meaningful texts from meaningless ones.

In perusing collections of inauthentic and authentic texts, we observed there seems to be a kind of information flow or semantic coherence that the authentic texts possessed, but the inauthentic did not. Given that this flow is a kind of pattern, we arrived at the conjecture that *there is a discernable and actionable correlation between meaning and compression in texts.* In Figure 1 we have given a stylized depiction of how we imagine compression and meaning are related with respect to the length-normalized expository texts. Given that the above-mentioned conjecture holds, we might be able to somehow exploit text compressibility to separate the two kinds of text.

In this paper we use supervised learning to distinguish between authentic texts and several classes of inauthentic texts. Our method is based on employing the universal source coding algorithms to generate features in a high-dimensional space, then applying support vector machines to discriminate between the classes. Our results indicate this is indeed possible for a number of different kinds of inauthentic texts. Based on the compression profiles we observed, the authentic texts were bounded by various classes of more compressible or less compressible texts. First, these results support our conjecture that there *is* a relationship between meaning and compressibility or, in other words, some nonlinear connection between subjective and objective (as defined by Hartley [5]) measure of information. Second, our results indicate that the learning algorithm based upon the compression profile outperforms standard text categorization schemes on several non-trivial classes of inauthentic texts. Last, on a deeper level, there seem to be subtle, long range patterns in meaningful texts that could prove elusive for computer generated texts to emulate.

2 Related Work

Distinguishing between authentic and inauthentic technical documents is a binary text categorization problem. In such a setting, one of the many text categorization schemes could be followed [16], but arguably the best performance has been achieved using word-based data representations, *e.g.* bag-of-words or TF-IDF, and subsequently employing supervised learning algorithms [6]. There are several reasons, however, why the word-based data representation may fail to do well for this task, at least as the only model. First, it is insensitive to the order of words, and it can be easily argued that a random permutation of words or sentences within an authentic document can make such documents indistinguishable. Second, standard text classification models are inherently category-specific and frequently rely on the identification of just several words to be successful. On the contrary, authentic papers can generally belong to any category of documents and are defined by semantics, not necessarily by word frequencies or word co-occurrence. Third, documents for which a large corpus cannot be easily collected may not be accurately predicted due to small dataset sizes. For such categories, it would be important to use "semantic models" from other authentic papers that could help in the categorization task.

Source coding algorithms have already found respectable application in various areas of text mining, as recently predicted by Witten *et al.* [15]. They represent a viable approach in cases when context needs to be incorporated into the classification scheme. For example, since lossless compression algorithms are capable of finding frequently occurring patterns in text, these approaches can generally be trained on a corpus of related documents to produce a code that can later be exploited to detect other similar documents. Such a scheme was recently adopted by Frank *et al.* [3] who attempted to classify text into various categories based on the PPM compression algorithm [2]. Similarly to this approach, Khmelev and Teahan [7] defined a repetition-based measure that could be used for text categorization and plagiarism detection. Other areas where compression schemes were employed include authorship attribution [8], extraction of generic entities, token segmentation and acronym extraction [14, 15].

Another popular computational approach to discerning the meaning of words as they appear in a particular context is Latent Semantic Analysis (LSA), which addresses the problems of synonymy and polysemy to find related words to a context of document. A detailed description of LSA and an overview of its various practical applications is provided by Landauer *et al.* [9].

3 Methods

A collection of 1,396 authentic papers, $\mathcal{A}$, written in English, was manually collected from several on-line journal archives accessible to our institution. In cases of some journals, only a couple of articles were freely available to us, while for the others there were no limits imposed. Our goal was to make a reasonably diverse set of authentic documents according to topic, style of exposition, and length. We included 32 different journal collections, and added another one, Other, consisting of randomly picked scientific papers from researchers' homepages.

In addition to $\mathcal{A}$, we collected 1,000 inauthentic English documents, which were obtained by querying *SCIgen - An Automatic CS Paper Generator* (`http://pdos.csail.mit.edu/scigen/`). We refer to this collection as $\mathcal{I}_{MIT}$. Since *SCIgen* provides only a limited diversity of inauthentic texts we generated several other classes of such documents in order to present our predictor with increasingly more difficult classification problems. The following types of documents were generated: (*i*) documents obtained as per-character permutations of authentic documents, (*ii*) documents obtained as per-word permutations of authentic documents, (*iii*) documents obtained by concatenating blocks of b consecutive words extracted from various authentic papers, where block size $b \in \{2^i, i = 0..8\}$, (*iv*) repetitive documents obtained by concatenating the same block of b consecutive words extracted from authentic papers, where $b \in \{2^i, i = 0..8\}$, (*v*) documents obtained by concatenating variable-length blocks from the authentic documents, where the length of each block was randomly selected from $\{2^i, i = 0..8\}$. These five classes of inauthentic documents are subsequently referred to as $\mathcal{I}_{char}$, $\mathcal{I}_{word}$, $\mathcal{I}^b_{block}$, $\mathcal{I}^b_{rep}$, and $\mathcal{I}^v_{block}$, respectively. All classes of $\mathcal{I}$, jointly refered to as $\mathcal{I}_{all}$, were obtained from the preprocessed authentic documents in order to mimic possible automated procedures of generating inauthentic texts. The length of each inauthentic paper was equal to the length of a randomly chosen authentic paper.

To preprocess data, each paper was first read into a string variable and all capital letters were lowercased. All non-letter characters were then replaced by a white space (words separated into two lines by a dash were rejoined) and the string was split into an array based on space delimiting. Next, all words shorter than 2 characters and longer than 20 characters were removed, as well as the stop words. Porter stemming [11] was then performed on the remaining words that were subsequently rejoined into a string and separated by a single white space.

To construct features from the preprocessed documents we employed the Lempel-Ziv [17] and Bender-Wolf [1] algorithms. Each document was compressed using a set of 10 sliding windows $w \in \{2^i, i = 1..10\}$, after which 20 compression factors were calculated [12]. The rationale for such feature choice comes from the fact that different windows, when used for document compression, have the ability to capture different long range string repetitions. Thus, combining them all together created a set of 20 numerical features, or a *compression profile*, that was subsequently fed into the supervised classification algorithm.

The classifiers were built using support vector machines, which are machine learning algorithms trained to maximize the margin of separation between positive and negative examples [13]. Given that the goal of our study was not to optimize prediction accuracy *per se*, but rather to provide a proof of concept, we have only used polynomial kernels with degree $p \in \{1, 2\}$. Thus, both linear and non-linear models were evaluated. We used SVM^{light} software and its default setting for the regularization parameter C [6]. Prior to SVM learning, the dataset was normalized using standard z-score normalization.

All predictors were evaluated using 10-fold cross-validation. We estimated sensitivity (true positive rate), specificity (true negative rate) and balanced-sample accuracy, defined as $acc = \frac{sn+sp}{2}$, where sn and sp are sensitivity and specificity, respectively.

4 Experiments and Results

We start with an analysis of compression performance on the various types of document classes. We initially compared the compression factor for the authentic texts versus various classes of inauthentic texts as a function of the window size used in compression. These comparisons indicate that the compressibility of authentic texts lies somewhere in-between various classes of inauthentic texts. In particular, on one side of the spectrum lie compression factors of classes $\mathcal{I}_{block}$, while on the other side of the spectrum are the repetitive texts from class $\mathcal{I}_{rep}$. Typical compression factors for the $\mathcal{I}_{block}$ class ranged between 0.97 (slight expansion) and 2.1. On the other hand, the $\mathcal{I}_{rep}$ class had wide variability of the compression factors, in some cases exceeding 1000 (*e.g.* for $\mathcal{I}^1_{rep}$). In general, the compression factors for both Lempel-Ziv and Bender-Wolf variants behaved according to our expectations.

Next, we ran extensive experiments and estimated prediction accuracy when authentic texts were discriminated against various classes of inauthentic texts. Interestingly, it proved to be very easy to discriminate against the $\mathcal{I}_{MIT}$ class, even for the very small window sizes used to create features. On the other hand, the behavior of the prediction algorithms was what we expected for various other classes of inauthentic documents. For example, discriminating between classes $\mathcal{A}$ and $\mathcal{I}^1_{block}$ was easy even for the small window sizes, while discriminating between $\mathcal{A}$ and $\mathcal{I}^{16}_{block}$, and $\mathcal{A}$ and $\mathcal{I}^{256}_{block}$ became more difficult for short windows, but still easy for the longer ones. Clearly, this behavior was an artifact of the way $\mathcal{I}_{block}$ classes were generated, but shows that the experimental results agreed with our expectations.

Initially, we were surprised by the prediction accuracy of 74% between $\mathcal{A}$ and $\mathcal{I}_{MIT}$ when the compression window was only $w = 2$. This accuracy increased to nearly 100% with an increase in window size, which proved that neither short range and especially not long range pattern repetitions resembled those of the authentic texts. We also investigated the situation in which class $\mathcal{A}$ was distinguishable from $\mathcal{I}^1_{block}$, even for $w = 2$. We found that there was a much larger spread of compression factors for the authentic texts as compared to $\mathcal{I}^1_{block}$. For example, the minimum observed compression factor for the class $\mathcal{I}^1_{block}$ was 0.916, while the maximum was .924. Thus, the statistical properties of $\mathcal{I}^1_{block}$ were fairly stable, which does not hold true for the authentic texts (0.909 and 0.941 respectively). An accuracy achieved by the linear support vector machine between $\mathcal{A}$ and $\mathcal{I}^1_{block}$ was 54%, but increasing the degree of the polynomial to $p = 2$ lead to an accuracy of 81%. On the other hand, a comparison between $\mathcal{A}$ and $\mathcal{I}_{MIT}$ reveals similar spreads of compression factors, but there was a shift in the average compressibility which caused these classes to be distinguishable even for the small window sizes.

Class of $\mathcal{I}$	TF$_{p=1}$	TF$_{p=2}$	CP$_{p=1}$	CP$_{p=2}$
$\mathcal{I}_{MIT}$	100%	100%	99.8%	99.7%
$\mathcal{I}_{char}$	N/A	N/A	100%	100%
$\mathcal{I}_{word}$	48.4%	49.6%	96.5%	95.8%
$\mathcal{I}^2_{rep}$	73.7%	70.0%	100%	100%
$\mathcal{I}^{16}_{rep}$	70.5%	66.7%	100%	100%
$\mathcal{I}^{256}_{rep}$	72.1%	66.0%	60.3%	80.4%
$\mathcal{I}^1_{block}$	64.9%	65.6%	100%	100%
$\mathcal{I}^{16}_{block}$	65.4%	65.0%	100%	99.9%
$\mathcal{I}^{256}_{block}$	72.0%	73.7%	92.0%	89.6%
$\mathcal{I}^v_{block}$	70.8%	69.5%	97.8%	96.8%
$\mathcal{I}_{all}$	N/A	N/A	67.1%	68.5%

Table 1: The accuracy of the predictor when distinguishing authentic papers from various forms of inauthentic papers. TF represents the term frequency feature set, and CP represents compression profile feature set.

Table 1 shows the classification accuracy of our system for various classes of inauthentic texts and four different learning procedures. We used support vector machines of degrees $p = 1$ and $p = 2$ for both term-frequency and compression-profile feature sets. While the compression based approach was very fast and used only 20 features, the system based on the term frequency contained nearly 200,000 features and took significantly more time to execute. In order to train these classifiers we used feature selection filters based on the statistical tests followed by the principal component analysis (PCA). For the feature selection filters we employed standard t-test with the p-value threshold of 0.01 to eleminate the feature, while the dimensionality after the PCA was kept at 30 without further experiments.

We also performed cross-validation experiment per type of journal, *i.e.* 33-fold cross-validation where each journal was the test set, in order to explore the influence of the style and vocabulary used in a particular journal on the performance accuracy. However, this problem has also proved to be of similar difficulty for the compression based system. Furthermore, we also explored the effects of length with truncated authentic papers (to length 100, 200, 400 and 800 words) versus $\mathcal{I}_{MIT}$. In these experiments, we used only a reduced set of features to avoid boundary effects, but with a similar outcome to the overall accuracy.

It is worth commenting that training all authentic *vs.* all inauthentic texts resulted in a somewhat lower prediction accuracy (Table 1), similarly to the observed outcomes in the study by Frank *et al.* [3]. Increasing the accuracy of the overall system requires more sophisticated model selection procedures and will be subject of our future work.

5 Discussion

In this paper we described a topic-independent supervised learning approach that is applied to the problem of discriminating between authentic and various classes of inauthentic technical (or expository) texts. We required that inauthentic documents be machine generated in order to explore differences between human generated informative text and various possible classes of inauthentic text. Our system is based on features generated by the two related compression algorithms, Lempel-Ziv and Bender-Wolf. However, in order to account for the various long range pattern repeats, a set of cascading window sizes was used to run the software. This approach could easily be extended to a much longer list of publicly and commercially available packages in order to help to detect inauthentic documents.

Due to the huge variability within the class of inauthentic papers, the detection of the authentic texts can be considered to be a one-class problem. This could lead to an application of various outlier detection methods that have been extensively studied in the literature. However, for the purposes of this paper and studying whether authentic texts are separable from the various classes of inauthentic texts, we believe that it was sufficient to use supervised approaches as their performance can be quantified in well-characterized ways. One could easily think of various combined models that could detect not only inauthentic documents, but at the same time be topic specific too.

In general, identifying meaning in the technical document is difficult, and we do not claim herein that we have found a way to distinguish between meaning and nonsense. We only emphasize that there are many nontrivial classes of inauthentic documents that can be easily distinguished based on compression algorithms. A likely reason for this is that the authentic documents possess some latent long-range word co-occurrences and pattern repeats, which contribute to the flow of the paper and are a necessary condition to convey meaning in a technical document. Naturally, there might exist a class of expository texts that may not fully, or at all, convey any meaning, but still possess the flow resembling those of the authentic documents. However, it is unclear to us how these could be generated especially if they had to fulfill complex language models and adhere to a topic too.

Our initial conjecture that meaning and compression are related in informative texts raises intriguing questions about other kinds of high level structures, *e.g.* between the so-called *blog* and non-blog, or between various categories of articles such as entertainment or news. In any case, we believe that the topics touched upon in this study, those related to the interplay between information, meaning and compressibility, warrant further investigation.

References

[1] P. Bender and J. Wolf. New asymptotic bounds and improvements on the lempel-ziv data compression algorithm. *IEEE Transactions on Information Theory*, 37(3):721–729, 1991.

[2] J. G. Cleary and I. H. Witten. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, COM-32(4):396–402, April 1984.

[3] E. Frank, C. Chui, and I. H. Witten. Text categorization using compression models. In J. A. Storer and M. Cohn, editors, *Proceedings of DCC-00, IEEE Data Compression Conference*, pages 200–209, Snowbird, US, 2000. IEEE Computer Society Press, Los Alamitos, US.

[4] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW2005*, pages 235–312, May 10-14 2005.

[5] R. V. L. Hartley. Transmission of information. *Bell System Technical Journal*, 7:535–563, 1928.

[6] T. Joachims. *Learning to classify text using support vector machines.* Kluwer, New Jersey, 2002.

[7] D. V. Khmelev and W. J. Teahan. A repetition based measure for verification of text collections and for text categorization. In *SIGIR '03*, 2003.

[8] O. Kukushkina, A. Polikarpov, and D. V. Khmelev. Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2):172–184, 2001.

[9] T. Landauer, P. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[10] G. Miller and N. Chomsky. Finitary models of language users. In R. Luce, R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, New York, 1963. Wiley.

[11] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[12] D. Solomon. *Data Compression: The Complete Reference. 2^{nd} ed.* Springer-Verlag, New York, NY, 2000.

[13] V. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, New York, NY, 1998.

[14] I. Witten. Applications of lossless compression in adaptive text mining. *Conference on Information Sciences and Systems*, Princeton University, 2000.

[15] I. H. Witten, Z. Bray, M. Mahoui, and W. J. Teahan. Text mining: A new frontier for lossless compression. In *Data Compression Conference*, pages 198–207, 1999.

[16] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90, 1999.

[17] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, IT-23(3):337–343, 1977.