Using Statistical Properties to Enhance Text Categorization

Rached ZANTOUT
Electrical and Computer Engineering Department, Rafik Hariri University
Beirut, Lebanon

and

Ziad OSMAN
Electrical Engineering Department, Beirut Arab University
Beirut, Lebanon

ABSTRACT

Statistical properties extracted from text are useful in many areas. Knowing who authored some text or knowing the category of a text is among the uses of collecting such statistics. In this paper, language-independent properties of text are studied using two categorized corpora of news articles. It is observed that the properties do not depend on the corpus nor on its size. Several interesting properties are identified which enable minimizing the training set for an intelligent categorization system. Aside from text categorization, the properties can be used to compare the information content between different corpora. The properties can also be used to compare the rate of new information content between different corpora.

Keywords: Statistical Properties, text categorization, text mining, data mining.

1. INTRODUCTION

Statistical properties of text such as word count and N-Grams have been studied thoroughly in order to enhance Natural Language Processing (NLP) operations. One of the main applications of NLP is text categorization. Knowing the language of a text or knowing to which categorical hierarchy a text belongs, are two applications of text categorization. Humans have a remarkable ability to categorize text in predetermined categories. It is believed that humans rely on language and text properties to be able to do such a categorization. Among such properties, ancient Arabs used character frequency for cryptographic purposes [1]. In [2], Arabic entropy was studied. It was found that the first order entropy was 4.21, the second order entropy was 3.77 and the third order entropy was 2.49. The corpus used in the study was 60 newspaper articles which contained 64,000 characters spread between 10,897 words. Those entropies make Arabic more redundant than other languages. The ability to classify text into categories is useful in many applications [3]. A computer which is able to do such a classification would be able to retrieve information faster and more efficiently, identify topics better and filter texts in a more efficient way. Even searching and sorting files can then be done faster and more efficiently.

In section 2 the literature is reviewed to introduce similar work. In section 3 the methodology used to collect the statistics from the corpora is described. The actual numbers are presented using tables and figures. In section 4, the point of separation from linearity is presented as a tool to help researchers decide whether their corpus size is enough. In section 5 the results of sections 3 and 4 are analyzed to explore the importance of distinct blocks, their numbers as well as the point of separation from linearity. In section 6, the paper is concluded with a summary of the work and suggestions for future research.

2. LITERATURE REVIEW

The most important concept in being able to classify text is the ability to quantify the similarity between two different texts. In [4] the theory of distance between texts was developed, it was then used in [3], [5] and [6] to be able to identify the author of an article based on a corpus of articles of known authors. An accuracy of 90% was achieved using decision trees to classify Arabic text in [7]. The properties of the corpus had an effect on the accuracy. Decision trees, when complemented with Support vector machines (SVM), K-Nearest Neighbor, and Naïve Bayes classifiers had better performance as reported in [8, 9 and 10]. Naïve Bayes classifiers alone resulted in an average accuracy of 62.7% as reported in [11]. Association rules alone had an accuracy of 74.41% in [12]. Developing their own distance-based classifier, authors in [13] reported 62% recall and 74% precision. Light stemming helped statistical methods raise their accuracy to 98% in [14]. However, stemming deteriorated the results for text categorization in [15]. SVM alone produced an F-measure of 88.11 in [16]. However, in [1] SVM was reported to give an accuracy of 68.65% while C5.0 was reported to give an accuracy of 78.42%.

In [17] three known stemmers namely Khoja, Light Stemmer, and n-Gram were used with the Naive Bayesian algorithm to classify texts. A Macro F1 average of classification of 0.83 was reported. A hybrid approach was used in [18] that produced superior results to those reported in [17]. Stemming was used as a pre-processing step to breakdown words into roots and stems. In [19], a feature reduction method was proposed. This method was used in [20] to enhance the effectiveness of Neural Networks and Support Vector Machines in Arabic text classification. Neural networks were reported to be superior to SVM. Dictionary-lookup stemming was found to be better than root-based stemming and light-stemming in conjunction with ANN classifiers. Moreover, it was reported that for SVM classifiers, light stemming is better than root based stemming and dictionary-lookup stemming methods.

WordNet was used in [21] for document classification. The multivariate chi-square test was used to reduce dimensionality. Using WordNet ameliorated the macro-averaged F1 value. In [22], The Bag-of-Words (BOW) was used with the Bag-of-Concepts (BOC) to perform text categorization. The corpus used was a collection of Wikipedia articles. In [23], Wordnet was used to discover concepts in texts. Those concepts were used with SVM, Decision trees, and kNearest Neighbors to classify texts. Including concepts was proven to produce better results on two different corpora especially for SVM.

3. STATISTICAL PROPERTIES OF TEXT

Two large corpora (BBC and CNN) were used in order to get properties of text [24]. Each corpus consisted of articles classified in five categories (Business, Entertainment, Science, Sports, World News). Table 3.1 shows the number of articles for each category. All articles in each category were processed to produce a list of distinct (none repeated) blocks (dblocks) in the category. Blocks were chosen to be either 4 or 6 or 8 bytes long.

Table 3.2 shows the results for category Science for both BBC and CNN and also the percentage of the ratio dblocks/blocks for that category. It is clear that by increasing the block size (4, 6, 8 Byte) the number and relative percentage of dblocks increase.

Table 3.1 Number of articles per category

	Business	Entert.	Science	Sports	W.News
BBC	296	122	232	219	1,489
CNN	836	474	526	762	1,010

Table 3.2 Blocks vs. Dblocks for the Science category

Blocks vs. Dblocks	BBC SCIENCE	CNN SCIENCE
Blocks	985,344	2, 205,952
DBLOCKS (4 Bytes)	10,697 (1.08%)	18,615 (0.84%) 1.299180e+04
DBLOCKS (6 Bytes)	56,965 (5.78%)	92,520 (4.19%) 6.668773e+04
DBLOCKS (8 Bytes)	154,685 (15.7%)	1.834589e+05 279,422 (12.67%)

Figures 3.1 through 3.10 represent graphs of blocks vs. dblocks for all categories in both BBC and CNN corpora for different size blocks. The curve for larger number of bytes per block is always above the curve for smaller number of bytes per blocks. This means that the number of distinct blocks increases with the size of the block. In all graphs also, the increase in the number of distinct blocks (Dblocks) will start to taper off when more blocks are introduced to the corpus, because of the language redundancy. This is logical since the more a human writes the more they will use the same words especially prepositions and proper names. The number of new words and the number of new sentence structures will not vary much between different articles especially if the person is writing a lot of articles. Tables 3.3 and 3.4 show the blocks versus the dblocks for the rest of the categories with different block sizes of 8 bytes, 6 bytes and 4 bytes. It is clear that the number of the dblocks and their percentage with respect to the blocks go down with the decrease in block size as seen previously in table 3.1.

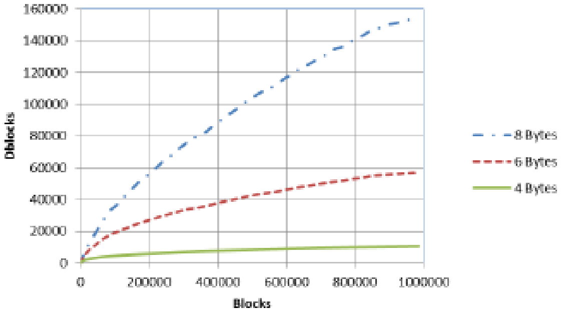

Figure 3.1 BBC Science blocks vs. dblocks for different block sizes.

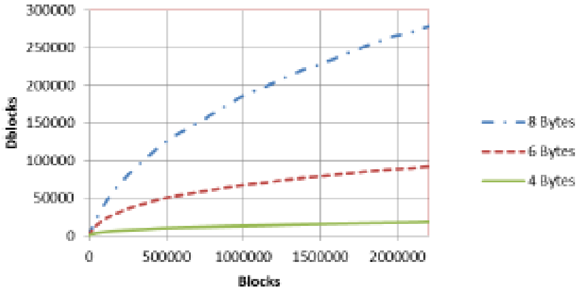

Figure 3.2 CNN Science blocks vs. dblocks for different block sizes.

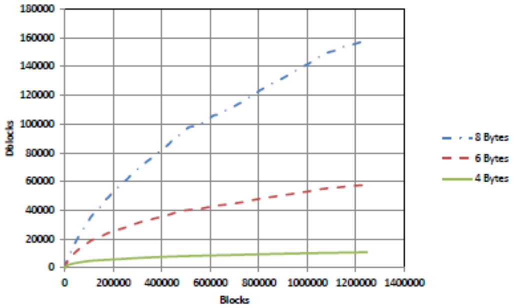

Figure 3.3 BBC Business blocks vs. dblocks for different block sizes.

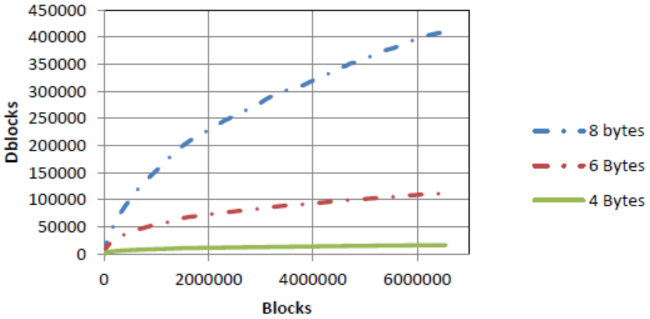

Figure 3.4 BBC World News blocks vs. dblocks for different block sizes.

Going below 4 bytes in block size is not beneficial because, even though this will result in lower number of dblocks, the distinction between categories will not be possible. Such a behavior is evident from both corpora which indicate that this property is independent of the corpus and is rather a linguistic property.

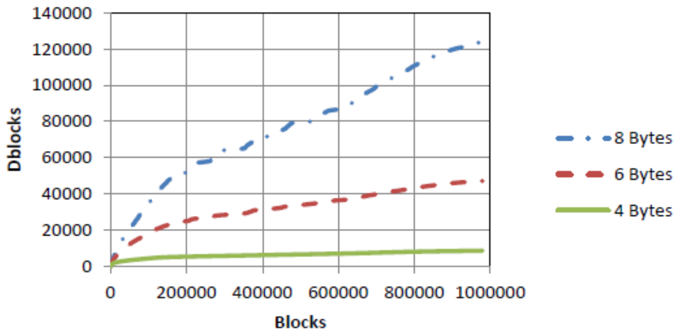

Figure 3.5 BBC Sports blocks vs. dblocks for different block sizes.

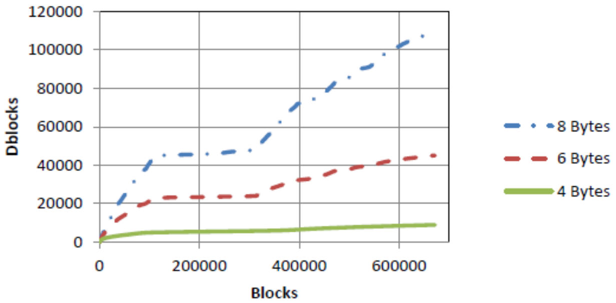

Figure 3.6 BBC Entertainment blocks vs. dblocks for different block sizes.

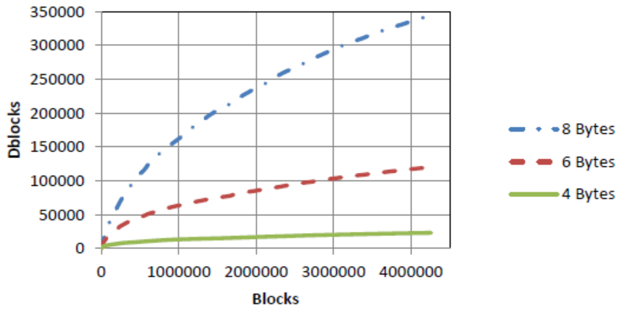

Figure 3.7 CNN Business blocks vs. dblocks for different block sizes.

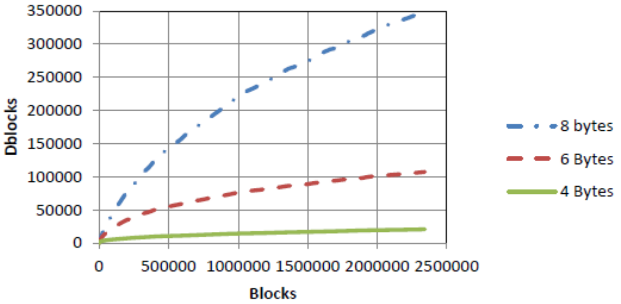

Figure 3.8 CNN Entertainment blocks vs. dblocks for different block sizes.

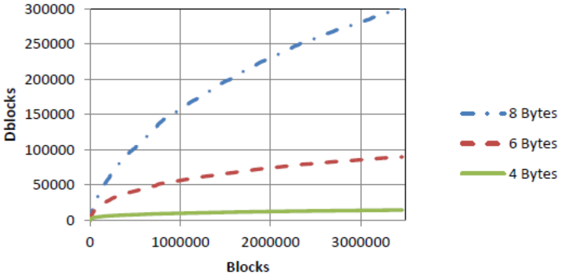

Figure 3.9 CNN Sports blocks vs. dblocks for different block sizes.

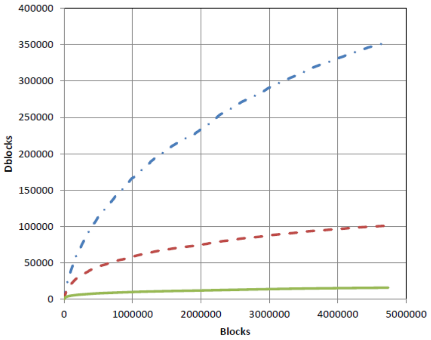

Figure 3.10 CNN World News blocks vs. dblocks for different block sizes.

Table 3.3 Blocks vs. Dblocks for Business and Entertainment categories

Blocks vs. Dblocks	BBC BUSINESS	CNN BUSINESS	BBC ENTER.	CNN ENTER.
BLOCKS	1,246,177	4,260,399	671,531	2,342,238
DBLOCKS (4 Bytes)	10,857 (0.87%)	23,000 (0.53%)	8,992 (1.33%)	20,606 (0.88%)
DBLOCKS (6 Bytes)	57,730 (4.63%)	120,015 (2.81%)	45,036 (6.7%)	107,106 (4.57%)
DBLOCKS (8 Bytes)	157,952 (12.67%)	345,379 (8.1%) 1.847e+05	109,345 (16.28%)	349,232 (14.19%) 1.712709e+05

Table 3.4 Blocks vs. Dblocks for the Sports and World News categories

Blocks vs. Dblocks	BBC SPORTS	CNN SPORTS	BBC WORLD	CNN WORLD
BLOCKS	979,308	3,460,922	6,543,954	4,736,957
DBLOCKS (4 Bytes)	8,569 (0.87%)	14,345 (0.41%)	16,520 (0.35%)	15,838 (0.24%)
DBLOCKS (6 Bytes)	47,021 (4.8%)	89,716 (2.59%)	111,780 (1.71%)	101,109 (1.54%)
DBLOCKS (8 Bytes)	123,784 (12.64%)	300,397 (8.68%) 1.565e+05	409,742 (6.26%) 3.499e+05	353,562 (5.4%)

4. POINT OF SEPERATION FROM LINEARITY

The point of separation from linearity (POSFL) is a measure of when the blocks added to the corpus will start to saturate the dblocks. This point is calculated by first finding the 9th degree polynomial equation fit for the blocks versus dblocks and then differentiating the equation. This will result in a graph similar to the one shown in Figure 4.1. This figure resembles that of the transfer function of a low pass filter. A horizontal line at 0.707 of the maximum vertical value will cross the curve at the point of separation from linearity.

The POSFL defines where the number of new blocks starts to taper off which means that not much new information is being added. This means that the articles before the point of separation are sufficiently representative of the category and can be used to train a system to classify new text.

Table 4.1 shows that the POSFLs for the Science category. One can notice that the number of dblocks is proportional to the number of bytes in a block.

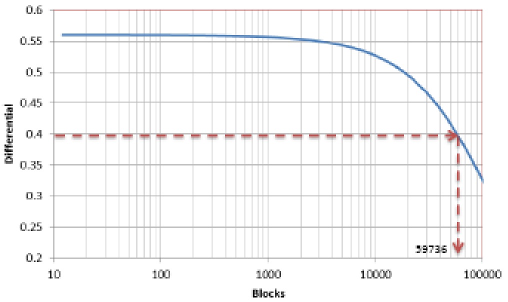

Figure 4.1 CNN Science POSFL determination.

Table 4.1 POSFL for different block sizes for the Science category

POSFL	BBC SCIENCE	CNN SCIENCE
BLOCKS	985,344	2,205,952
DBLOCKS (4 Bytes)	5845, 0.59%	25,008, 1.13%
DBLOCKS (6 Bytes)	17,844, 1.81%	36,065, 1.63%
DBLOCKS (8 Bytes)	27,643, 2.8%	59,736, 2.7%

Figures 4.2 through 4.11 show the differential of the 9th degree polynomial fit of the dblocks versus blocks data for all categories for different block sizes. It is noticeable that they all have the same shape as a low pass filter. This means that the POSFL is independent of the category or the source of the articles.

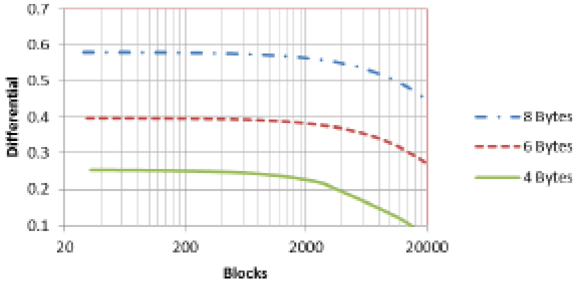

Figure 4.2 BBC Science differential for different block sizes

Tables 4.2 and 4.3 show a similar trend as seen in table 4.1, for the other categories.

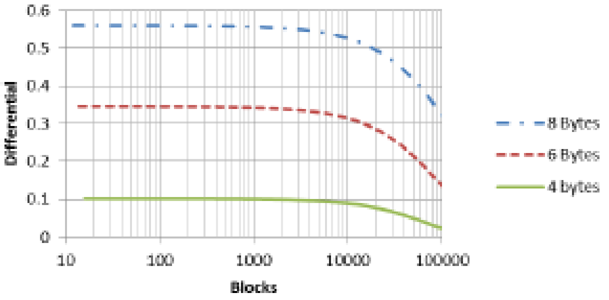

Figure 4.3 CNN Science differential for different block sizes.

The change in the blocks size has an effect on the position of the point of separation from linearity and on the percentage of distinct blocks. The trend in figures 4.12 and 4.13 is increasing with the increase of the block size. This is logical since as the block size increases, the variety in the blocks will increase. This is because the number of combinations between the bytes in a block increases as the size of the block increases.

Table 4.2 POSFL for different block sizes for the Business and Entertainment categories

POSFL	BBC BUS	CNN BUS	BBC ENTER.	CNN ENTER.
BLOCKS	1,246,177	4,260,399	671,531	2,342,238
DBLOCKS (4 Bytes)	6,098, 0.49%	47,940, 1.125%	1,879, 0.28%	27,981, 1.19%
DBLOCKS (6 Bytes)	24,082, 1.932%	65,499, 1.53%	11,226, 1.67%	40,951, 1.75%
DBLOCKS (8 Bytes)	42,978, 3.44%	101,160, 2.37%	14,838, 2.21	79,026, 3.73%

Table 4.3 POSFL for different block sizes for the Sports and World News categories

POSFL	BBC SPORTS	CNN SPORTS	BBC WORLD	CNN WORLD
BLOCKS	979,308	3,460,922	6,543,954	4,736,957
DBLOCKS (4 Bytes)	2,168, 0.22%	28,998, 0.83%	60,623, 0.92%	38,998, 0.82%
DBLOCKS (6 Bytes)	18,181, 1.85%	46,913, 1.35%	84,534, 1.29%	63,003, 1.33%
DBLOCKS (8 Bytes)	50,814, 5.19%	67,051, 1.93%	12,2710, 1.87%	88,173, 1.86%

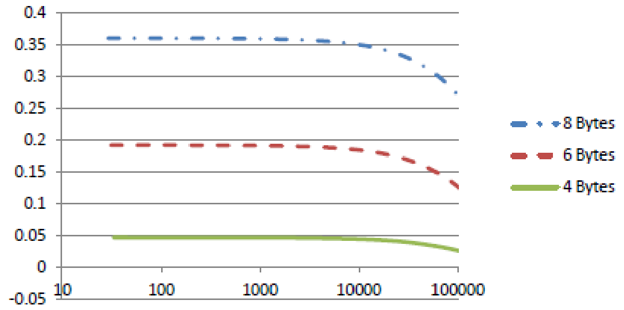

Figure 4.4 BBC World News differential for different block sizes.

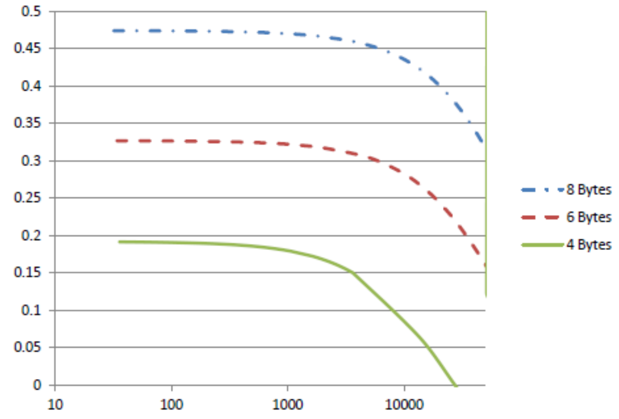

Figure 4.5 BBC Business differential for different block sizes.

Figures 4.12 through 4.13 show the POSFL trend for BBC and CNN for different categories versus block size.

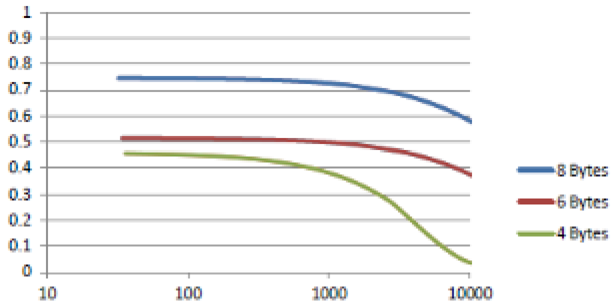

Figure 4.6 BBC Entertainment differential for different block sizes.

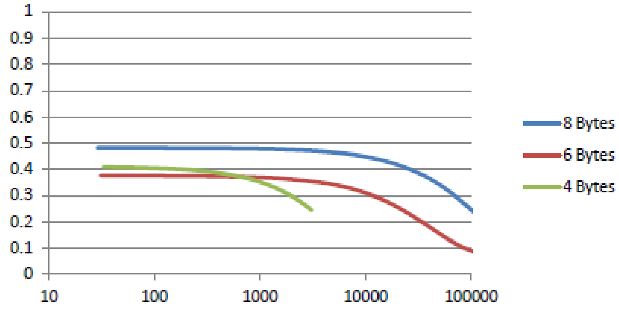

Figure 4.7 BBC Sports differential for different block sizes.

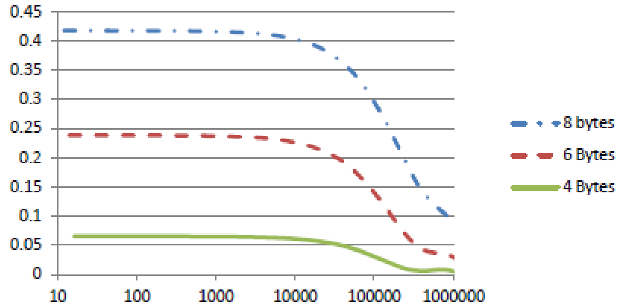

Figure 4.8 CNN Business differential for different block sizes.

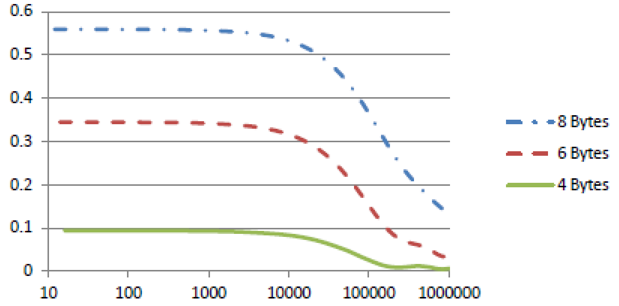

Figure 4.9 CNN Entertainment differential for different block sizes.

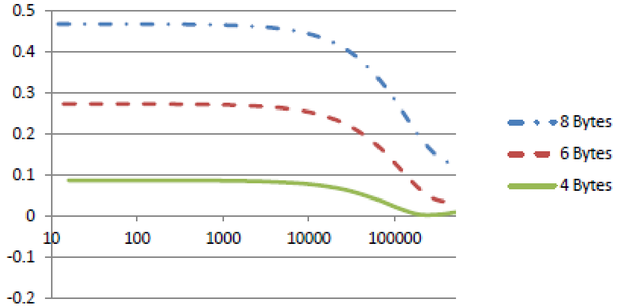

Figure 4.10 CNN Sports differential for different block sizes.

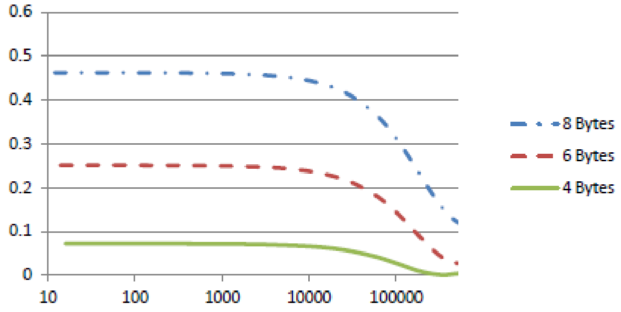

Figure 4.11 CNN Worlds News differential for different block sizes.

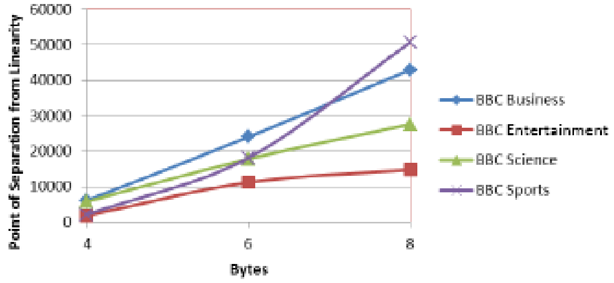

Figure 4.12 BBC categories POSFL for different block sizes.

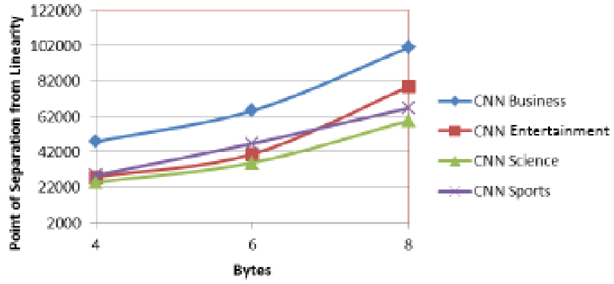

Figure 4.13 CNN categories POSFL for different block sizes

5. ANALYSIS

When choosing the BBC and CNN Science blocks to be of the same value (in this case the lower value of 985,344 blocks of BBC Science) , as shown in table 5.1, the numbers of dblocks for CNN Science is always higher than those for BBC Science regardless of the block size.

This also evident in Figure 5.1 which shows the trend of dblocks versus blocks for the same number of blocks (985,344 blocks) for BBC and CNN Science using 8 byte block size. The CNN Science curve is always above that of the BBC Science curve for all block values. In fact this is true for all categories as shown in Figures 5.2 through 5.5.

Table 5.1 Blocks vs. Dblocks for the Science category having the same number of blocks

Blocks vs. Dblocks	BBC SCIENCE	CNN SCIENCE
Blocks	985,344	985,344
DBLOCKS (4 Bytes)	10,697	12,991
DBLOCKS (6 Bytes)	56,965	66,687
DBLOCKS (8 Bytes)	154,685	183,458

Table 5.2 Blocks vs. dblocks for the Business and Entertainment categories having the same number of blocks

Blocks vs. Dblocks	BBC BUSINESS	CNN BUSINESS	BBC ENTER.	CNN ENTER.
BLOCKS	1,246,177	1,246,177	671,531	671,531
DBLOCKS (8 Bytes)	157,952	184,741	109,345	171,275

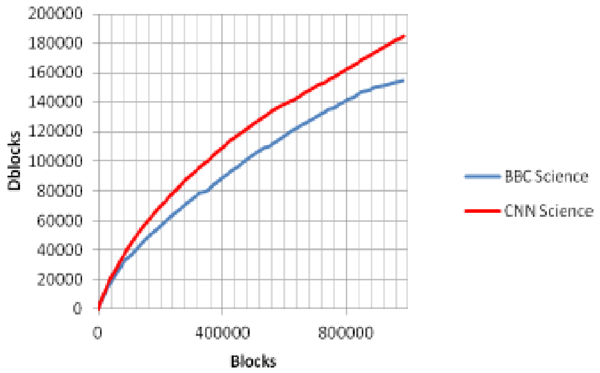

Figure 5.1 BBC and CNN Science blocks vs. dblocks for 8 byte block size

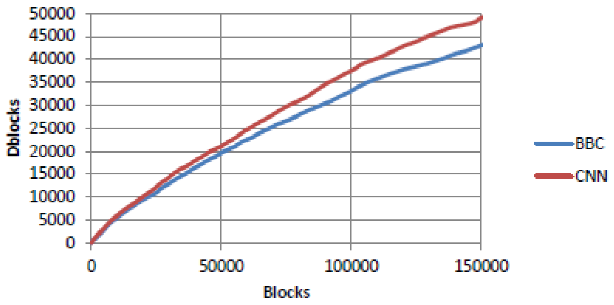

Figure 5.2 BBC and CNN Business blocks vs. dblocks for 8 byte block size

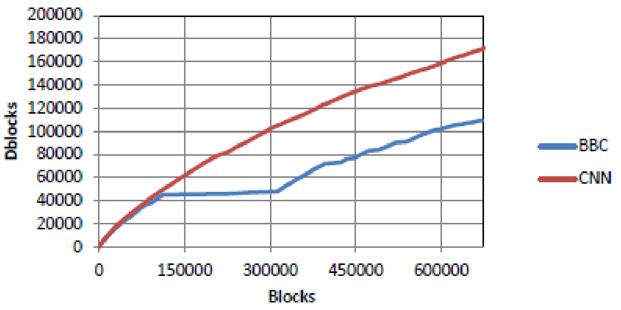

Figure 5.3 BBC and CNN Entertainment blocks vs. dblocks for 8 byte block size

Choosing the lower value of blocks as a common value for BBC and CNN, as shown in tables 5.2 and 5.3 for the 8 byte block size, it is noticed that the same trend exists.

Even when the BBC blocks are originally more than the CNN blocks like in the case of the World News category (6,543,954 versus 4,736,957) CNN has always a higher number of dblocks compared to a similar sized corpus from BBC. This shows in Figure 5.5 for both BBC and CNN having the same number of blocks, in this case it is 4,736,957 blocks versus dblocks with 8 byte block size. The CNN curve is always higher than that of the BBC curve.

Table 5.3 Blocks vs. dblocks for the Sports and World News categories having the same number of blocks

Blocks vs. Dblocks	BBC SPORTS	CNN SPORTS	BBC WORLD	CNN WORLD
BLOCKS	979,308	979,308	4,736,957	4,736,957
DBLOCKS (8 Bytes)	123,784	156,522	349,978	353,562

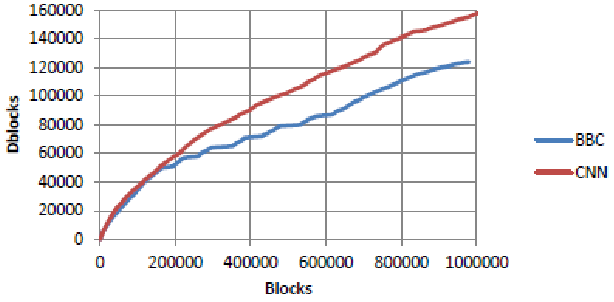

Figure 5.4 BBC and CNN Sports blocks vs. dblocks for 8 byte block size

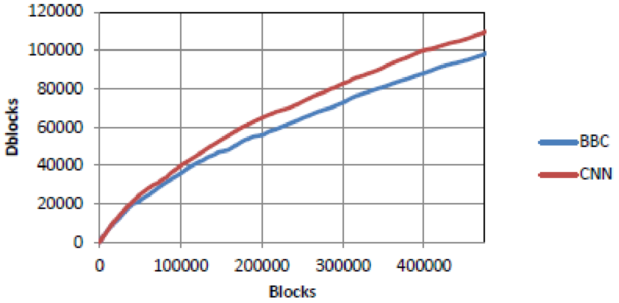

Figure 5.5 BBC and CNN World News blocks vs. dblocks for 8 byte block size

This means that CNN articles introduce more new information per article than does BBC. This could also mean that writers in CNN have a wider variety in their writing styles compared to writers in BBC.

6. CONCLUSIONS

In this paper statistical properties of text were proven to possess several interesting characteristics. It was observed that there was a direct relationship between the number of distinct blocks in text belonging to a category and the point of separation from linearity. It was also observed that a small fraction of the blocks that are contained in text in a specific category is representative of the whole category. This means that only a small fraction of the blocks in a text are required to train a system to classify text correctly. It was also observed that different corpora can be compared to find out which one has more new information using the distinct blocks. It was also observed that as the size of the block increases, the number of distinct blocks increases also.

Future work should focus on quantifying the above observations with actual analysis of the text to prove the conclusions. Work should also be done on corpora from different languages to study the effect of the entropy of the language on the distinct blocks and point of separation from linearity. Furthermore, an effort should be done to check the distinct blocks of each category and compare distinct blocks between categories. It would be interesting to determine the percentage of blocks which are specific to a category and how this percentage grows with the size of the category. The distinct blocks of each category should be used in a text categorization system to compare the performance of a system which used distinct blocks with the performance of more traditional systems. Finally, it would be interesting to see whether the above conclusions and observations are still valid if, instead of using blocks of bytes, blocks of words are used.

9. REFERENCES

[1] Simon Singh, **the Code Book**, 1999, pp 14-20.

[2] Wanas, M.; Zayed, A.; Shaker, M.; Taha, E. “first-second- and third-order entropies of Arabic text”, **IEEE Transactions on Information Theory**, Volume 22, Issue 1, Jan 1976, Page: 123.

[3] Ayadi, Mohsen Maraoui, Mounir Zrigui, **Intertextual distance for Arabic texts classification**, Rami IEEE 09.

[4] Muller C “**Principes et methodes de statistique lexical**” Paris, Hachette universite, 1977

[5] Thomas M. “An Application of Authorship Attribution by Intertextual Distance in English” **Corpus Numero 2, La Distance intertextuelle**, December 2003

[6] Labbe C. and Labbe D “Inter-Textual Distance and Authorship Attribution. Corneille and Moliere” **Journal of Quantitative Linguistics**, 8-3, December 2001, pp. 213-231.

[7] Fouzi Harrag, Eyas El-Qawasmeh, Pit Pichappan, “Improving Arabic Text Categorization using Decision Trees’, **IEEE 09**.

[8] H Sawaf, J. Zaplo, and H. Ney “Statistical Classification methods for Arabic news articles” **Workshop on Arabic Natural Language Processing, ACL01**, Toulouse France, July 2001.

[9] F. Sebastiani “Machine Learning in Automated Text categorization” **ACM Computing Surveys**, 2002, 34(1) pp. 1-47.

[10] Y. Yang and J. O. Pedersen “A comparative Study on feature selection in text categorization” **Proceedings of the 22nd ACM International Conference on Research and Development in Information retrieval**, SIGIR 99, ACM Press, NY, USA, 1999, pp. 42-49.

[11] M. El-Kourdi, A Bensaid, and T. Rachidi “Automatic Arabic Document Categorization Based on Naïve Bayes Algorithm” **20th International Conference on Computational Linguistics**, Geneva, August 2004.

[12] A. El-Halees, “Arabic Text Classification Using Maximum Entropy” **The Islamic University Journal (Series of Natural Studies and Engineering),** 2007, 15 (1), pp. 157-167.

[13] R. M. Duwairi. “ A distance-based classifier for Arabic text categorization” In **Proceedings of the International Conference on Data Mining**, Las Vegas, USA, 2005.

[14] M. M. Syiam, ZT, Fayed and M. B. Habib, “An Intelligent System for Arabic Text Categorization” **IJICIS**, 2006, 6(1), pp. 1-19.

[15] H., Froud, R. Benslimane, A. Lachkar, A. Lachkar, S. Alaoui Ouatik, “Stemming and Similarity Measures for Arabic Documents Clustering”, **IEEE 2010**.

[16] A. A. Mesleh, “Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System” **Journal of Computer Science**, 2007, 3(6), pp. 430-435.

[17] M.Hadni, A.Lachkar, and S. AlaouiOuatik, “A New and Efficient Stemming Technique for Arabic Text Categorization", **IEEE, 2012**, pp. 791-796.

[18] M. Hadni, S. AlaouiOuatik and A. Lachkar, "Effective Arabic Stemming Based Hybrid Approach For Arabic Text Categorization", **International Journal of Data Mining & Knowledge Management**, Vol.3, No.4, 2013, p1.

[19] FouziHarrag, Eyas El-Qawasmah and Abdul Malik S.Al-Salman, "Stemming as a Feature Reduction Technique for Arabic Text Categorization", **IEEE, 2011**, pp. 128-133.

[20] M. Ikonomakis, S.Kotsiantis, V. Tampakas “Text Classification Using Machine Learning Techniques” **WSEAS Transactions On Computers**, Issue 8, Volume 4, 2005, pp. 966-974.

[21] Alaa Alahmadi, Arash Joorabchi, and Abdulhussain E. Mahdi, "Combining Bag-of-Words and Bag-of-Concepts Representations for Arabic Text Classification", **ISSC 2014 / CIICT 2014**, Limerick, June, 2014, 26-27.

[22] H. Nezreg, H. Lehbab, and H. Belbachir, "Conceptual Representation Using WordNet for Text Categorization", **International Journal of Computer and Communication Engineering**, Vol. 3, No. 1, January 2014.

[23] Zakaria El berrichi and KarimaAlbidi," Arabic Text Categorization: A comparative Study of different representation Modes", **The International Arab Journal of Information Technology**, Vol. 9, No.5, September 2012, pp.465-470.

[24] Motaz K. Saad and Wesam Ashour, "OSAC: Open Source Arabic Corpus", **6th ArchEng International Symposiums, EEECS’10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science**, European University of Lefke, Cyprus, 2010.