

5.4.2 USENET Postings

The USENET postings were taken from the Westbury Lab USENET corpus (2005-2010) [31]. This data was chosen as it is freely accessible and the language used in the text is from normal internet users and so is a good test for the algorithms effectiveness for "web language" text where there may be many spelling and grammar mistakes. There are three different postings totalling 616 words, with any usernames or other personal information removed. The topics are email clients and the naming of villains in fiction.

Results

	Bitstream 1	Bitstream 2
Words	616	616
Total Bitrate	64	63
Average bitrate (/10 words)	1.03	1.02
False Positives	11	16
False Negatives	1	0
True Positives	66	59
True Negatives	538	541
Error Count	0	0
Original Total Quality	579641	579641
Original Average Quality	940	940
Obfuscated Total Quality	684284	647641
Obfuscated Average Quality	1110	1051

Table 5.2: Results for Usenet Postings

Analysis

This text file is perhaps the most successful of all of the files processed. It has achieved a good bitrate (just over 1 bit per 10 words) and also has achieved an improvement in the quality.

The false positive rate is consistent with the other text files, and it does experience a single false negative with the first bitstream. There were no errors found in the results. Interestingly, even though the bitrate was only very slightly lower for the second bitstream, the number of false positives is considerably higher. This will be due to replacement words failing deobfuscation checks with the available bit, but the original word being deobfuscatable.

5.4.3 Academic Paper Extract

The academic paper is the introduction section from Ross Anderson's paper "Information Hiding - A Survey" [27]. This text contains reference to some technical terms and names