SUMMARY

In this thesis we have discussed the problem of text classification by machine learning, and an innovative tool to be used in solving that problem. The core of the work done by the author has been to devise an algorithm to interpret vectorized text according to the principles and properties of a Gaussian distribution. The product of this work has been a method of defining the two key parameters of such a Gaussian distribution: mean and covariance, such that they can be interacted with in a non-Euclidean, angular inner product space. While in principle any of the tools applicable to a gaussian distribution can be applied by extrapolation from these definitions, the author has gone on to explicitly define one tool: Mahalanobis distance, by just such an extrapolation.

In order to adequately interpret and interact with text written in human natural language, a machine must have a sophisticated, and organic method of interpreting that text. A machine must be able to extract not just concrete data, but "meaning" from text. In a year of precursor work to this thesis, the author has investigated systems understanding text with just such an objective: the word2Vec and RIV systems. These systems, which produce human understandable relationships between words, provide a fascinating bridge between the concrete machine data form of text, and the abstract intuition of human speech. In the course of the authors study of these systems, a pattern of Gaussian properties seemed to arise in the vectors so produced. However, it was found that the existing tools to interact with such properties do not interface smoothly within the measurement systems of angular distance which are integral to understanding these vectors.

For this reason, an examination of these properties and the necessary methods of interfacing with them has been undertaken. The product of initial investigation was that the interface would have to provide an inner product rule definition which could accommodate these properties in the angular space in which text classification occurs.

From this starting point, an inner product rule was so defined, which relies on the Euclidean properties of the tangent distance function to interface between any angular distance function and Gaussian properties. Additionally, covariance, mean, and Mahalanobis distance were detailed in terms of this inner product rule.

After the mathematical system was described, an implementation was tested, yielding very encouraging results. Although this implementation was not the core focus of this thesis, and this analysis is not exhaustive, it begs further investigation in the coming months.