2017-9

“How Short is a Piece of String?”: An Investigation into the Impact of Text Length on Short-Text Classification Accuracy

Austin McCartney
Technological University Dublin

“How short is a piece of string?”: An Investigation into the Impact of Text Length on Short-text Classification Accuracy

Austin McCartney

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

2017

DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signature Date: June 25th 2017

Austin McCartney

ABSTRACT

The recent increase in the widespread use of short messages, for example micro-blogs or SMS communications, has created an opportunity to harvest a vast amount of information through machine-based classification. However, traditional classification methods have failed to produce accuracies comparable to those obtained from similar classification of longer texts. Several approaches have been employed to extend traditional methods to overcome this problem, including the enhancement of the original texts through the construction of associations with external data enrichment sources, ranging from thesauri and semantic nets such as Wordnet, to pre-built online taxonomies such as Wikipedia. Other avenues of investigation have used more formal extensions such as Latent Semantic Analysis (LSA) to extend or replace the more basic, traditional, methods better suited to classification of longer texts.

This work examines the changes in classification accuracy of a small selection of classification methods using a variety of enhancement methods, as target text length decreases. The experimental data used is a corpus of micro-blog (twitter) posts obtained from the 'Sentiment140'[1] sentiment classification and analysis project run by Stanford University and described by Go, Bhayani and Huang (2009), which has been split into sub-corpora differentiated by text length.

Key words: text classification, short text, naïve-Bayes, support vector machine, latent semantic analysis, twitter, enrichment, enhancement.

[1] http://help.sentiment140.com/for-students/

ACKNOWLEDGEMENTS

I'd like to sincerely thank my supervisor, Dr. Svetlana Hensman, and dissertation co-ordinator, Dr. Luca Longo, for their input and support during the course of the creation of the initial proposal and execution of this work.

I'd also like to thank Mr. Bill Parker and Mr. Stacey Tarro for their generous flexibility in facilitating me in my completion of this project.

Finally, I'd like to thank my family for their patience, support and encouragement.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

1.1. Research Focus

The focus of this research is to characterise the performance of binary classification of short texts as a function of the target text length, when using a variety of text enhancement methods in conjunction with a small selection of classifiers.

1.2. Background

Traditional supervised learning techniques for machine classification of texts, for example a bag-of-words approach using a naïve-Bayes classifier, rely on statistical methods which in turn rely on a sufficiency of 'meaningful' data, (words), within the texts to allow differentiation into classes. In the case of short texts, the performance of such classifiers is reported as being poor in comparison with their performance on longer texts, because insufficient data is present within the body of the target texts.

Despite quite extensive coverage in published literature of the general area of short test classification, very little specific information has been available relating to the deterioration of classifier performance for shorter texts; the exact nature of the relationship between text length and classifier performance has been unclear and, consequently, no common definition of how short a target text may be before it can be considered troublesome is available.

One promising avenue for the improvement of classifier performance has been the enhancement of the short text by the addition of synonyms, or other semantically linked words, to the body of the original text prior to classification. As its simplest, this consists of adding synonyms from a thesaurus for all words present in a text, but more sophisticated methods can use the text to 'concept mine' related terms from a semantic net such as Wordnet[2], or from a pre-built classification scheme such as those used to organise encyclopaedias like Wikipedia[3]. The implicit hope in such supplementation of the text is that the additional words are conceptually related to the words in the original text and will therefore 'amplify' the underlying meaning and context of the original.

[2] http://wordnet.princeton.edu/

[3] https://en.wikipedia.org/wiki/Wikipedia:FAQ/Categories

1.3. Research Project/Problem

The primary focus of this work is defined by the research question:

"Do the changes in performance of different short-text classification methods, as measured by weighted average accuracy (F1 Score), differ between text enhancement methods and classifiers as target text length decreases?".

In attempting to answer this question, the experimental portion of this project will investigate two related sub-questions:

1.3.1 Research Sub-question 1

"Does classification performance for any of the included text enhancement methods change as texts decrease in length?"

In graphical terms, this may be thought of as asking whether the plot of classification performance against text length for any enhancement method has a significant non-zero slope. For example, Enhancement A, in Figure 1-1 below, shows no change in performance with decreasing text length, while Enhancements B and C appear to show some change in performance with respect to text length. In practical terms, any method showing robust performance with respect to decreasing text length, such as Enhancement A below, would substantially 'solve' the difficulties associated with short text classification.

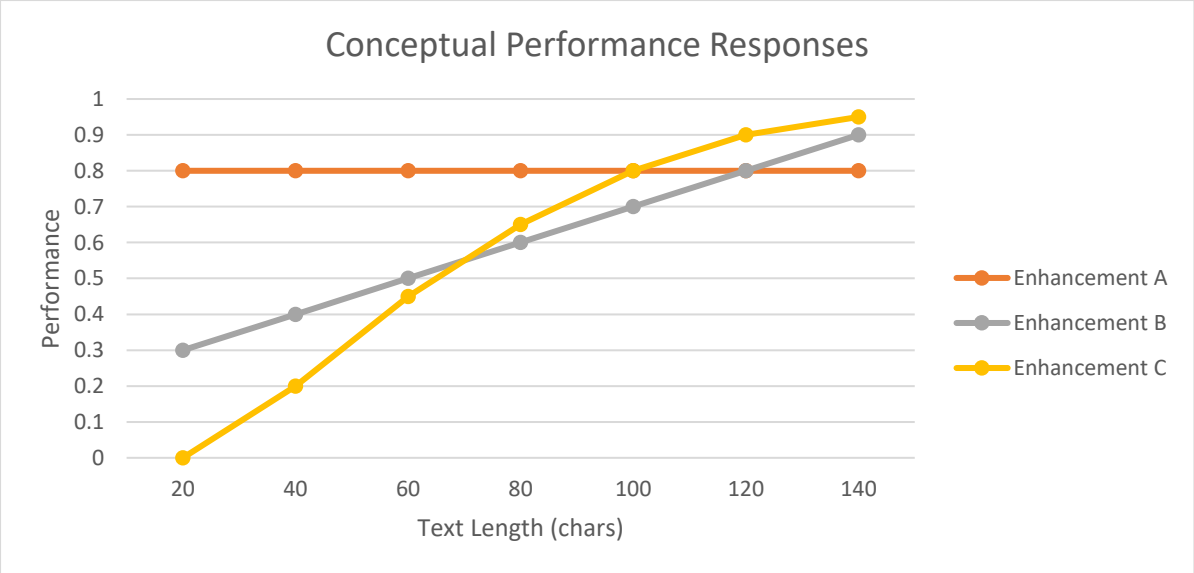

Figure 1-1 Conceptualizing changes in classifier performance with decreasing text length

1.3.2 Research Sub-question 2

"If classification performance varies, as texts decrease in length, for two or more of the included text enhancement methods, do those performance variations differ between enhancement methods?"

In graphical terms, this can be understood as asking whether there is a significant difference between the curves plotted for Enhancements B and C in Figure 1-1 above. In practical terms, any method showing superior performance with respect to decreasing text length, such as Enhancement B above, which shows more robust behaviour than C, would, all else being equal, be a preferred method for short text classification.

These sub-questions will be more formally stated as experimental hypotheses in the next section.

1.4. Research Objectives

Two main hypotheses will be tested during the course of this work. These null hypotheses are, in order of specificity:

1.4.1 Hypothesis 1

The performance of short-text classification enhancement methods, as measured by weighted average accuracy (F1 Score), will not change as target text length decreases.

1.4.2 Hypothesis 2

The changes in performance of different short-text classification enhancement methods as target text length decreases, measured by weighted average accuracy (F1 Score), will not differ between enhancement methods.

1.4.3 Research Objective 1

Measure and analyse the performance of selected classification enhancement methods with respect to message length.

1.4.4 Research Objective 2

Measure and analyse the changes in relative performance of selected classification enhancement methods with respect to message length.

1.4.5 Experimental Tasks

The high-level tasks undertaken to achieve the research objectives were:

- Obtain and prepare master dataset
- Build enhanced datasets using selected enhancement methods
- Split data sets into subsets according to message length
- Train and test classifiers for different message length subsets and methods
- Measure accuracy within enhancement methods with respect to message length
- Test for significant differences in text-length related reduction in accuracy between classification enhancement methods

1.5. Research Methodologies

The research associated with this project will be secondary and quantitative, as it relies on a corpus of data previously collected by the Stanford University's Sentiment140 project and will carry out a series of binary text classification experiments upon subsets of that data with a view to conducting statistical analyses to determine the character of the inter-relationships between the text length of messages within the corpus and the accuracy of the binary classification of those messages. The nature of the statistical analyses and the chosen hypotheses dictate that the research will be empirical rather than theoretical. This work has, primarily, inductive characteristics as it will attempt to verify the commonly held view that text classification becomes more difficult as text length decreases, and will also try to establish general characteristic behaviours of a range of text enhancement methods based on experimental results.

This work is not intended to create a production software solution and, as a result, the commonly used CRISP-DM[4] methodology is not completely applicable throughout its lifecycle. However, the broad outline of the early stages of the CRISP-DM model will be followed. The Data Understanding phase of CRISP-DM will be covered in Section 4.1. The Data Preparation phase corresponds to section 4.2 of this work. Section 4.3 is indicative of the iteration between CRISP-DM's data Understanding and Modelling phases, and comprised some pre-experimentation and adjustment before modelling began in earnest. Section 4.4 will cover the Modelling phase, and the evaluation phase corresponds to the remainder of chapter

[4] http://www.comp.dit.ie/btierney/BSI/CRISP-DM%20Process%20Model.pdf

4 and the entirety of chapter 5. In the context of this work CRISP-DM's Business Understanding phase may be considered analogous to the Literature Review at the start of the cycle, and at the end of the CRISP-DM cycle, to the Conclusions which are outlined in Chapter 6.

1.6. Scope and Limitations

The scope of this work is strictly limited to the examination of the changes in classification performance as text length decreases and comparison of the performance of text enhancement methods when used in conjunction with a particular classifier. Specifically, no attempt beyond the most basic was made to optimise or tune classifier performance, and any reference to the comparative performance of classifiers is made in an informal sense. The use of multiple classifiers was undertaken only in order to demonstrate the general applicability of the findings, if any, and to rule out any effect that may arise from the use of any specific classifier: reflecting this purpose, the three classifiers chosen were used in their most basic configurations.

It should also be noted that the experimental twitter data from the Sentiment140[5] project originally formed part of that project's training data, and that the classes assigned by that project, used as this project's training and test data, were determined based on the presence of either positive or negative emoticons within the body of the 'tweet'. Positive emoticons were assigned a positive sentiment and negative emoticons were assigned a negative sentiment. It is possible, since all the pre-classified tweets, by design, contained either a negative or positive emoticon, that these messages may not be representative of short messages in general in terms of the ease of classification with respect to sentiment: it would not seem unreasonable to conjecture that a tweet-author motivated to add an emoticon to a message may be writing a message containing more significant sentiment than the average message. If true, this might imply that the task of differentiating between positive and negative sentiment in the Sentiment140 corpus will be a relatively easy classification task.

1.7. Document Outline

The remainder of this document will be laid out as follows: Chapter 2 will review the available published literature relating to the classification of short texts, methods of enhancing short texts prior to classification and commonly used classifiers within the short-text domain. The

[5] http://help.sentiment140.com/for-students/

Literature Review corresponds to the initial Business Understanding phase of the CRISP-DM model.

Chapter 3 will discuss the underlying design of the experiments used in this study, including the statistical treatments of experimental results.

Chapter 4 will cover the implementation of the experimental design including data acquisition, data preparation, enhancement of tweet texts and classification steps. Chapter 4 will also present the results of experiments in both graphical and tabular form. This Chapter corresponds to the Data Understanding, Data Preparation and Modelling phases of the CRISP-DM lifecycle.

Chapter 5 will discuss statistical testing of results and draw quantitative conclusions on the specific hypotheses set out in Section 1.3 above. This Chapter corresponds to the Evaluation phase of the CRISP-DM lifecycle.

Chapter 6 will conclude the main body of this work, and evaluate the design, results and conclusions presented in earlier chapters, and discuss gaps and opportunities for refinements and further work. The Conclusion represents the re-visiting phase of CRISP-DM's Business Understanding phase.

Chapter 7 will comprise a bibliography of relevant published work, presented in APA6 referencing format.

Chapter 8 will be formed of appendices relating to additional relevant information including a glossary of terms.

2. LITERATURE REVIEW AND RELATED WORK

2.1. Supervised Text Classification

King, Feng, and Sutherland (1995), in their wide-ranging comparative study, outline several general cautions relating to supervised classification which can be assumed to continue to hold today: specifically, they observe that classifier performance is very dependent on the particular data set in use, that studies using 'only a handful' of classifiers are inherently limited and that classifiers need to be carefully tuned, preferably by an expert, before optimum performance can be realised. They also recommend that Bayes classifiers are not used for sets containing any significant degree of co-variation between variables. However, Lim, Loh and Shih (2000) found that there was little difference in accuracy over a large selection of classification algorithms used on a selection of real world datasets and suggest, in production situations, that training time requirements, scalability and comprehensibility of results should be given more weight in the algorithm selection process.

Holte (1993), in a paper relating specifically to one-level decision tree classifiers, makes the more general observation that simple problems often respond very well to simple classification approaches and characterises his datasets with the comment:

> The practical significance of this research was assessed by examining whether or not the datasets used in this study are representative of datasets that arise in practice. It was found that most of these datasets are typical of the data available in a commonly occurring class of 'real' classification problems. Very simple rules can be expected to perform well on most datasets in this class.

2.1.1 Naïve-Bayes

Lewis (1998) describes naïve-Bayes as “a favorite [sic] punching bag of new classification techniques” but this is only partially borne out in a general review of the more recent published work on this algorithm. It might be more accurate to say that modern research takes a sensibly nuanced approach to naïve-Bayes and recognises that within certain contexts it can perform on a par with far more sophisticated, and more costly, methods. In their work, specifically designed to update the work of King, Feng and Sutherland (1995), Caruana and Niculescu-Mizil (2006) agree with the earlier authors in their opinions on naïve-Bayes, but also qualify

this caution by commenting "These generalizations [on naïve-Bayes], however, do not always hold." They found that Support Vector Machines, on average, significantly out-performed naïve-Bayes but they conclude that there is significant variation across problems between methods. A different perspective was demonstrated by Peng and Schuurmans (2003) who augmented naïve-Bayes with n-gram techniques to good effect, in a method they named 'chain augmented naïve-Bayes', over a set of text classification problems and they conclude that their technique yields "state of the art performance that competes with the best-known methods in these cases" due to the inherent relaxation on the 'naïve' independence assumption.

Kim, Han, Rim and Myaeng (2006) make two points in relation to the text classification of naïve-Bayes which are particularly significant in the context of the current work. They propose that the perceived weaknesses of naïve-Bayes in the domain of text classification are due, firstly, to poor parameter estimation related to inaccurate estimation of word frequencies which is caused by an imbalance in document sizes. They comment "parameter estimation in this model is affected more by long documents than by short documents; the longer a document, the more terms participate in parameter estimation". In the context of this project, the document lengths for a given classification run are, by design, confined to a set of similar values. In a more general sense, "short" texts, such as tweets, may be considered to have broadly similar document lengths throughout the whole domain. Kim, Han, Rim and Myaeng's second identified weakness relates to an insufficient number of training samples for some categories: in the current work, this is patently not an issue as training sets were well populated and balanced between categories. The authors assert that by circumventing these weaknesses in their own experiments they demonstrated "our proposed naïve Bayes text classifier performs very well in the standard benchmark collections, competing with state-of-the-art text classifiers based on a highly complex learning methods such as SVM.".

Rennie, Shih, Teevan, and Karger (2003) identify imbalanced representations of target categories within training data as a primary weakness of naïve-Bayes and, once again, claim that the removal of this 'skewed data' weakness significantly improves classification performances.

The material absence of all of the above-mentioned weaknesses in the current work may contribute to an explanation of the comparatively successful performance of naïve-Bayes presented in Sections 4 and 5 below, which may appear unusual in light of so many caveats in the published literature.

2.1.2 Support Vector Machines

Although a more recent innovation than naïve-Bayes, Support Vector Machines (SVMs) have a long history of implementation in text classification problems. Early work by Cortes and Vapnik (1995) introduced SVMs specifically in the context of binary classification commenting particularly on their high capacity for generalization and their superior performance, when compared with what were then traditional methods, over a variety of problems.

Joachims (1998) also claimed substantial performance gains for SVM methods, and detailed reasons on why SVMs are particularly appropriate for text classification tasks. Joachim's four arguments were that 1) SVMs easily handled high dimensional space problems as the mechanisms for prevention of over fitting were not sensitive to the number of features. 2) The ability to handle high dimensional spaces is crucial since it can be shown that text has the property that even the less informative features still contain significant discriminatory information. 3) SVMs cope well with the sparse vectors that are characteristic of high dimensionality text problems. 4) Joachim asserts that "most text categorization problems are linearly separable" which facilitates the fitting of a hyperplane. Joachims also commented on the lack of a requirement for feature selection and concluded that the robustness of SVMs, the ability to run 'out-of-the-box' without parameter tweaking and their performance on text classification gave them a significant advantage over existing methods.

Support Vector Machine text classification, as related to sentiment analysis, is specifically addressed by Pang, Lee and Vaithyanathan (2002), who claim that sentiment analysis is a more difficult problem than standard document topic classification. They demonstrate that corpus-based machine learning models easily out-perform key-word list methods derived from human experience. They go on to claim that their SVM method out-performed a maximum entropy classifier, and that they increased its degree of advantage by binarizing their vectors to effectively remove word frequency information and rely only on binary word presence alone. Their further work with bigrams and part-of-speech tagging yielded no improvements in classification for any of the SVM, MaxEnt or naïve-Bayes classifiers that formed part of their experiments. They conclude with a discussion of some stylistic features of their target texts which they suspect may have increased problem difficulty beyond the difficulty of a simple topic classification including narrative techniques such as "thwarted expectations" which, they postulate, may create specific difficulties for bag-of-words approaches and which may have analogies in the related domain of short-texts.

Support Vector Machines in the context of sentiment analysis is also the specific focus of Mullen and Collier (2004) who augmented standard SVMs with part-of-speech techniques, semantic differentiation techniques and syntactic relations. They agree in their findings with Joachims (1998), that no variation from baseline SVM configuration was required, stating:

> Several kernel types, kernel parameters, and optimization parameters were investigated, but no appreciable and consistent benefits were gained by deviating from the default linear kernel with all parameter values set to their default.

2.1.3 Latent Semantic Analysis

The earliest corpus-based attempts to classify texts relied on 'surface matching' using frequency based methods such as Term Frequency-Inverse Document Frequency (TF-IDF), but these proved unsatisfactory for shorter texts due to the sparseness and brevity of the available data. Seminal work on Latent Semantic Analysis (LSA) conducted by Deerwester, Dumais, Furnas, Landauer and Harshman (1990) improved upon basic methods through the introduction of a dimensionality reduction step, implemented using Singular Value Decomposition, to enhance the relative information content of the vectorised texts. Work has continued on LSA in its original form by Landauer, Foltz and Laham (1998), and work has been carried out to extend and combine it with other methods such as Adaboost in the work of Cai and Hofmann (2003), with good results reported for longer texts, in an information retrieval (search) context.

A conceptual link between LSA, where "hidden semantic redundancies are tracked across (semantically homogeneous) documents" and n-grams where features and inter-relationships span only a range of a few words, is highlighted by Bellegarda (2000), who goes on to recommend a blend of LSA and n-grams as a promising future avenue of research. Bellegarda, however, also warns that bag-of-words techniques are insensitive to word order and gives an example where the placement of the word "not" in a text leads to a radical change of underlying meaning without a corresponding change in the vector representation of the text. This effect has obvious implications in the context of sentiment analysis where an error in the identification of the negated part of text could lead to a reversal of perceived sentiment.

Landauer, Laham, Rehder and Schreiner (1997) also discuss the inability of LSA to use word ordering, but conclude that it is not essential, and give an example of LSA at least matching

human performance on an experiment designed to test the ability to assess the quantity and quality of knowledge conveyed in short essays.

Another interesting insight into the underlying conceptual mechanisms of LSA is given by Kontostathis and Pottenger (2006) in which they discuss "higher-order term co-occurrence" as a mechanism inherent in the Singular Value Decomposition step of LSA: this co-occurrence is, at its root, the same mechanism by which the hypernym and hyponym text enhancements using wordnet, discussed below, hope to make use of in order to find and match underlying concepts embedded within the target texts and thereby allow establishment of relatedness to determine class. Kontostathis and Pottenger go on to observe, in the same article, that LSA is computationally very expensive, and that there is an extra parameter tuning step to be undertaken to determine the optimal truncation value for each dataset (see section 4.3 below). They suggest that future work might focus upon the search for "an algorithm for approximating LSI", and their subsequent discussion gives rise to some doubt as to the suitability of LSA for use in production systems due to these concerns.

2.2. Short Texts and Text Enhancement

A variety of different techniques have been proposed to enhance or enrich short texts by the addition of extra features designed to make matching, clustering and classification easier. Some of these rely on the exploitation of external taxonomies, typically Wikipedia or Probase, whereas others use semantic nets such as Wordnet. Others, such as Keller, Lapata and Ourioupina (2002) attempt to leverage the implicit information held by search engines through the development of similarity measures based on the frequency of term co-occurrences returned by search engines. Song, Ye, Du, Huang and Bie (2014) present a survey of short text classification, first giving an overview of the special conditions which attach to short text as a problem, and then outlining all the major avenues of current research. They divide approaches into three broad families – semantic approaches, including LSA, semi-supervised classical methods (e.g. SVM, naïve-Bayes) and ensemble methods, which can combine from the other two families.

Work was presented by Bollegala, Matsuo and Ishizuka (2007) which incorporated semantic information extracted from web-based search engines and this was contrasted with the same operation using Wordnet: the authors point out that, typically, a static resource such as Wordnet will fail to produce good results when trying to judge similarity in the presence of

colloquialisms. Hu, Sun, Zhang and Chua (2009) combine Wordnet feature enhancement with supplementation from the Wikipedia taxonomy using Wordnet for texts containing few non-stop-words and Wikipedia for texts containing many non-stop-words.

This use of an explicit external taxonomy such as Wordnet can be contrasted with much work which makes use of the implicit taxonomy inherent in the organisation and content of reference sources such as Wikipedia and Probase as in the work of Banerjee, Ramanathan, and Gupta (2007) where the titles of Wikipedia articles containing terms of interest were used as features to supplement the sparse text data, or in the work of Wang, Wang, Li, and Wen (2014) in which they coined the term 'bag-of-concepts' to stress the semantic aspect of the additional features that they had mined from the probabilistic semantic network Probase. Wikipedia is once again the favoured external source of 'world knowledge' in Gabrilovich and Markovitch (2006) in which they state, "pruning the inverted index (concept selection) is vital in eliminating noise", but, unfortunately, they provide no further detail on their 'ablation' process. Gabrilovich and Markovitch go on to claim double digit improvements over the then state-of-the-art methods on 'certain datasets'. A later paper by the same authors, Gabrilovich and Markovitch (2007), uses concepts mined from Wikipedia as the dimensions of a high dimensional concept space, and maps documents into this space as weighted vectors, which are then to compute semantic relatedness. They go on to contrast this technique, which they call Explicit Semantic Analysis, ESA, against latent semantic analysis, LSA, emphasising that the concepts in ESA are human generated concepts rather that statistical co-occurrences.

Genc, Sakamoto and Nickerson (2011) compared three disparate techniques to demonstrate the utility of Wikipedia as an implicit taxonomic source. In a manner similar to, but subtly different from, Gabrilovich and Markovitch (2007) they use the target text to mine relevant Wikipedia pages, and then calculate the distances between Wikipedia pages using a simple shortest path graph traversal metric to assign distances between target texts. Their second technique is to simply measure the String Edit Distance between texts using the Levenshtein metric. Their final design uses Latent Semantic Analysis coupled with a cosine distance metric. Their results suggest that the Wikipedia method out-performed both SED and LSA on most sets, and was inferior on none of the tested datasets.

In a change of tactic, Genc, Mason and Nickerson (2013), again using Wikipedia, have taken a different approach to concept extraction by using a sliding n-gram window centred on a target word within a text, and by using the explicit categorisation provided by Wikipedia category

containers which they have pre-classified into four super categories to provide meta-tags used to filter matching concepts.

Work by Agirre, Alfonseca, Hall, Kravalova, Pasca and Soroa (2009) has further extended the frequency-and-semantics based approach with the addition of syntactic context, achieved through the use of a variable window of text around key words enabling searching for a context specific word match, effectively implementing a sophisticated variable n-gram.

Departing from the common themes above, Sun (2012) takes a refreshingly contrarian direction to the main approaches outlined above, and trims short texts even further in an attempt to retain only key words. Trimming is accomplished using familiar term-frequency / inverse-document-frequency methods coupled with a novel 'clarity' measure, and is followed with a classification implemented through a Lucene search to find similar documents from a corpus: the classes of the returned documents are used as the class for document under classification. Sun reports that results match MaxEnt classifiers.

Sriram, Fuhry, Demir, Ferhatosmanoglu and Demirbas (2010) observe that:

> When external features from the world knowledge is [sic] used to enhance the feature set, complex algorithms are required to carefully prune overzealous features. These approaches eliminate the problem of data sparseness but create a new problem of the curse of dimensionality.

Rather than tackle those problems head-on, they exploit the available data relating to the authors of the short texts (tweets) to enhance the feature set of each method. Another paper taking a slightly different approach to the problem is Zelikovitz and Hirsh (2000) in which a second-order similarity relationship is used as a bridge: target texts are compared to a large unlabelled corpus of background knowledge, and matched corpus documents are then compared to labelled data. Despite this apparently distant relationship, Zelikovitz and Hirsh claim to have significantly reduced error rates although they do concede that the background corpus selection is critical to success, and that the combinatorial nature of the search may lead to efficiency and scaling problems.

A trend in the short text enhancement literature becomes apparent over time: early work concentrated on well-structured external resources such as Wordnet but, with time, the favoured approach became the less well-structured Wikipedia-type model. Several authors comment that Wikipedia's relatively wider domain provides better 'world knowledge' and can

therefore more effectively measure context and concept than can the relatively narrow-focus of Wordnet. Gabrilovich and Markovitch (2007) also place emphasis on the 'real knowledge' aspect of Wikipedia and explicitly prefer it to the latent nature of statistically derived concepts as used in Latent Semantic Analysis.

2.3. ANOVA and Robust Statistics

Simplystatistics.org[6] has referred to R.A. Fisher as 'the most influential scientist ever' in part due to his contribution of the workhorse statistical method of ANOVA, which is used to test for statistical differences between means. The ANOVA method makes some assumptions about the underlying structure of data and, although it has been often found to be robust with respect to violations of these assumptions, for example in Feir-Walsh and Toothaker (1974), more robust variations have been developed. One such family of robust methods, Wilcox Robust Statistics (WRS), has been developed by Wilcox and has not only been discussed in Wilcox and Keselman (2003), but has also been made available as a package for the R statistical environment[7]. Wilcox and Keselman state that:

> Conventional methods generally offer at most a small advantage in statistical power over modern methods when standard assumptions are approximately true. This is because modern methods are designed to perform nearly as well under these circumstances.

With this assurance in mind, Wilcox's trimmed means methods have been used for all AVOVA related tests in this work.

Other non-parametric tests of relevance to the current work are the Jonckheere-Terpstra test for trend detection, described by Jonckheere (1954) and Spearman's Rank Order Co-efficient, suitable for detecting correlation between two ranked sets, as described by Zar (1972).

2.4. Gaps

Although frequent reference is made to the difficulty of classifying short text, as for example in Song, Ye, Du, Huang and Bie (2014), all bar one of the reviewed articles omit any reference

[6] https://simplystatistics.org/2012/03/07/r-a-fisher-is-the-most-influential-scientist-ever/
[7] https://cran.r-project.org/web/packages/WRS2/index.html

to the quantitative impact of the shortness of the text or any definition of how short a text must be to be considered ‘short’. Yuan, Cong and Thalmann (2012) in their paper, which is concerned, primarily, with contrasting various smoothing methods as applied to naïve-Bayes, conclude only that classifiers perform more poorly with single word texts than with multi-word texts.

3. DESIGN / METHODOLOGY

This chapter will discuss the underlying design of the experiments used in this study, including the statistical treatments of experimental results.

The fundamental design of the experiment centres on measuring classification performance on enhanced variants of messages of known specific lengths when selecting messages into classes having either positive or negative sentiment.

The differences, if any, in classification performance across message lengths and between enhancement methods, as measured by the F1 score for accuracy of classification, will be analysed to determine if message length or enhancement has any statistically valid impact on classification performance, and so answer the research questions posed in section 1.3 above:

- "Does classification performance for any of the included text enhancement methods change as texts decrease in length?"
- "If classification performance varies, as texts decrease in length, for two or more of the included text enhancement methods, do those performance variations differ between enhancement methods?"

3.1. Sub-setting Data by Text Length

The original data set will be split into distinct message-length subsets, each subset containing 5000 tweets, all of exactly the same length when measured in characters, and having an even balance between tweets pre-categorised as having either positive or negative sentiment. There will be twelve length categories in total. The length categories, as measured by the total number of characters in the original message will be: 138 characters, 110 characters, 80 characters, 50 characters, 45 characters, 40 characters, 35 characters, 30 characters, 25 characters, 20 characters, 15 characters and a final set of tweets of length <= 10 characters.

The original experimental design called for only five length categories, of lengths 138 characters, 110 characters, 80 characters, 50 characters and 20 characters. However, early exploratory work, outlined in Chapter 4 below, suggested that a closer spacing of lengths would be advantageous.

3.2. Text-enhancement Data Preparation

Each tweet message in each of the length-determined subsets will be pre-treated with nine text enhancement techniques to produce a total of ten variants of each message, including the original message.

Broadly speaking, three approaches to enhancement will be used: basic, Wordnet-based and Wikipedia-based. A brief summary of each of the three approaches is given below: more detail on the exact nature of the enhancements will be included in the following chapter, and a full example of the results of all of the enhancements of a single tweet, taken from the Sentiment140 corpus, is presented in appendix 8.3.

3.2.1 Basic Enhancements

Basic enhancements consist of operations such as the removal of stop words, punctuation and twitter hashtags, the lemmatization of the text - the replacement of words with their simplest root form, and the creation of bigrams – strings comprised of consecutive word pairs occurring in the original text. (Bigrams are a specific 2-word instance of the more general *n*-gram concept which comprises strings of *n* consecutive words in a text).

3.2.2 Wordnet

Wordnet (Miller, 1995) is a semantically focused English language dictionary. It bears a resemblance to an extended thesaurus but, importantly from the perspective of this work, it contains not only synonyms, but also hypernyms, which are words representing related but less specific concepts than the searched word, and hyponyms, which are words representing related but more specific concepts than the searched word. Wordnet also provides some part-of-speech functionality. This project will make use of wordnet by appending the results of synonym, hypernym and hyponym searches to the cleaned & lemmatized version of the original tweet.

3.2.3 Wikipedia / DBpedia

DBpedia[8] is a static, structured, database derived from information contained in the online encyclopaedia Wikipedia. Importantly, from the perspective of this project, DBpedia provides a web-based interface which returns, in XML format, the Wikipedia taxonomic metadata for the most relevant Wikipedia pages when a given word or bigram is searched. This metadata includes page titles, Wikipedia categories and Wikipedia classes. These metadata each have a 'label' which is a text descriptor, possibly containing multiple words, of the page title, category

[8] http://wiki.dbpedia.org/

or class. For example, labels contained in the DBpedia metadata for the word 'dog' include 'animal' and 'mammal'. Wikipedia enhancements will be made by means of appending these metadata labels to the cleaned & lemmatized version of the original tweet. Appendix 8.2 presents, by way of example, the full DBpedia XML output for the single word "sound".

It may be noted that these three approaches to enhancement may be categorised into one of two classes: loosely speaking, the basic enhancements do not supplement the text with any external data if we discount the substitution of a word with its own lemma, whereas the Wordnet and Wikipedia/DBpedia approaches rely primarily on the addition of external data, which, it is implicitly hoped, is in some way conceptually linked to the words in the original text, thereby 'amplifying' the underlying meaning of the text.

3.3. Final Data Structure

Following the division by length of the original data, and the generation of enhanced message variants, the experimental data will consist of 120 separate data sets, each containing 2500 messages pre-classified as having positive sentiment and 2500 messages pre-classified as having negative sentiment. This can be conceptualised as shown by Figure 3-1 below, where each cell represents a data set of 5000 messages.

	Enhancement									
Text Length	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words
10										
15										
20										
25										
30										
35										
40										
45										
50										
80										
110										
138										

Figure 3-1 Conceptual View of Final Data Sets

3.4. Classification – Model & Test

Each of the 120 data sets will be used to repeatedly build and test classification models. This will be repeated 100 times for each of the 120 data sets. For each run of the modelling step the data will be randomly split into 90% training data, used to train a single model, and 10% test data used to test the classification performance of that model; this method is known as repeated random sub-sampling validation or Monte Carlo cross-validation (Xu & Liang, 2001). The F1 Score (weighted average of precision and recall) for each model will be recorded. This will result in 100 F1 Scores for each of the 120 data sets shown above. The graphical conception of this matrix of results can be seen in Figure 3-2 below, where the independent variable is plotted as text length and the dependent variable is the F1 score. Each data point represents the mean of the 100 F1 scores for each data set.

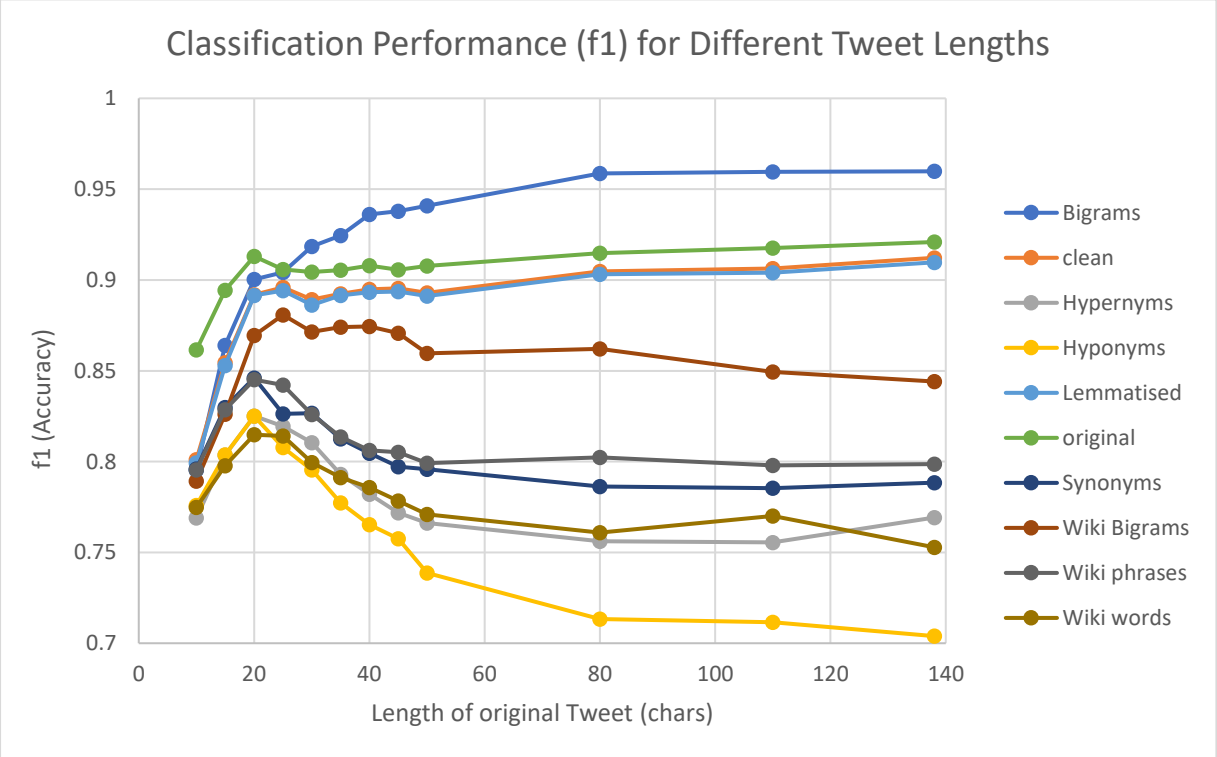

Figure 3-2 Sample Graph of Result Matrix for a Single Classifier

This procedure will be completed for each of three separate classifier methods as follows:

- Naïve-Bayes / Bag-of-Words
- Support Vector Machine (SVM)
- Latent Semantic Analysis followed by SVM

As discussed in the introduction to this work, the purpose of using multiple classifiers is to assess the general applicability of any findings rather than to formally assess the performance of the classifiers themselves, and so a detailed comparison of those classifiers' performance will not form a major part of this work.

3.5. Evaluation

Evaluation of the results for each of the classifiers will first be carried out using a 2-way independent analysis of variance, to determine whether statistical differences with respect to text length or enhancement are present.

2-way analysis of variance testing will be followed by application of 1-way analysis of variance for each of the enhancement method data sets and the non-parametric Jonckheere-Terpstra test for trend detection will be applied individually to each of the enhancement method data sets.

If significant differences in F1 Score are present with respect to text length, this will justify the rejection of Hypothesis 1: *The performance of short-text classification enhancement methods, as measured by weighted average accuracy (F1 Score), will not change as target text length decreases.*

The 2-way ANOVA test also includes an evaluation of the interaction effect between variables. This may be thought of as asking the question "Does the enhancement method have a significant effect on the way in which F1 Score changes with length?". If the interaction should prove to be significant this will justify the rejection of Hypothesis 2: *The changes in performance of different short-text classification enhancement methods as target text length decreases, measured by weighted average accuracy (F1 Score), will not differ between enhancement methods.*

In addition to the statistical testing of the hypotheses, 1-way analysis of variance testing and trend testing will be followed by application of post-hoc measures to assign both text-lengths and enhancement methods into significantly separate groupings with respect to classifier performance for all three classifiers in order to provide some descriptive results on the relative performance of enhancements and the characteristic behaviour of the classifiers as text length decreases.

3.6. Summary

The average performance of three classification methods will be tested on short texts of twelve different lengths, having 10 different enhancements applied to the texts before classification. The results of these tests will be statistically analysed to ascertain whether performance differs between applied enhancement methods or whether performance differs as text-length decreases. The next chapter will discuss the implementation and results of these experiments and chapter 6 will cover the statistical analysis of the results and the conclusions which may be drawn from that analysis which can be used to address the research question "*As target text length decreases, do the changes in performance of different short-text classification methods, as measured by average class accuracy, differ between methods?*"

4. IMPLEMENTATION / RESULTS

Chapter 4 will cover the implementation of the experimental design including data acquisition, data preparation, enhancement of tweet texts and classification steps. Chapter 4 will also present the results of experiments in both graphical and tabular form.

All required code for acquisition and enhancement of the data was written in Python 2.7[9], making extensive use of Python's NLTK[10] package, v3.2.2, for natural language processing functionality. The base classifiers used were those provided by Python's scikit-learn[11] package, v 0.18.1, and the version of Wordnet used was v3.0, which came bundled with NLTK. The code for the implementation of the Latent Semantic Analysis, the code for Wordnet enhancement and the code for Wikipedia enhancement were purpose built for this project in Python 2.7.

4.1. Data Acquisition and Inspection

The Sentiment140 tweet corpus was downloaded from http://help.sentiment140.com/home. The corpus consisted of 1.6 million tweets; 800,000 are pre-classified as having positive sentiment and 800,000 are pre-classified as having negative sentiment.

The large number of total available tweets allowed the selection of large, length specific, sub sets, each containing 5000 tweets. Figure 4-1 illustrates that at least 5000 tweets are available for all message lengths of, or in excess of, 15 characters.

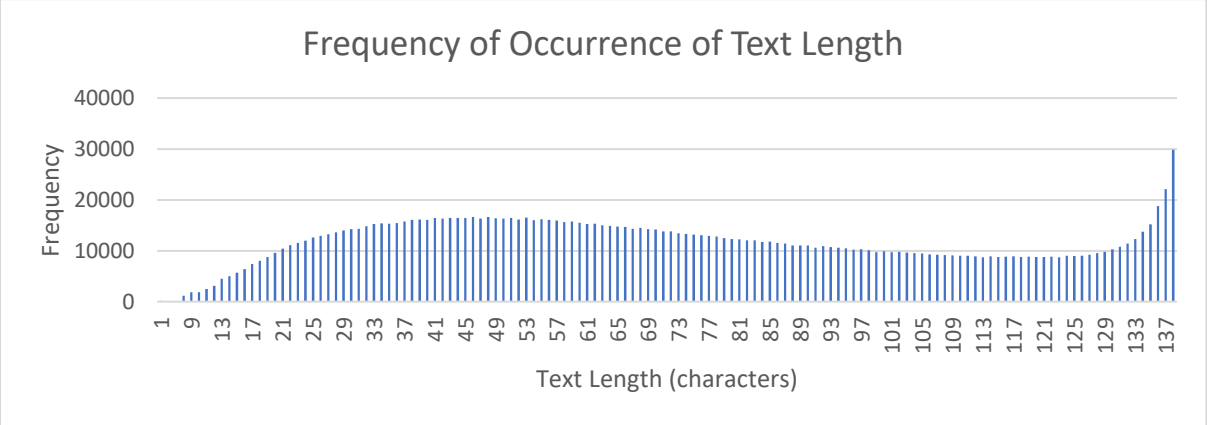

Figure 4-1 Histogram of Text Lengths in Sentiment140 Corpus

[9] https://www.python.org/download/releases/2.7/
[10] http://www.nltk.org/
[11] http://scikit-learn.org/stable/

From this total set, length-based subsets of 5000 tweets were randomly drawn (50% positive, 50% negative). The character lengths of the tweets in the respective subsets were exactly 138,110, 80, 50, 45, 40, 35, 30, 25, 20, 15 and a final set of tweets of length <= 10 characters.

It was notable, and convenient, that much basic data cleaning had already been carried out by the Sentiment140 project. The data did not contain unprintable control characters in the length ranges of interest. One consequence of this pre-cleaning was that the emoticon artefacts, for example ":)", had been stripped from tweets of the maximum length of 140 characters, leaving their actual length at 138 characters. For this reason, 138 characters was the maximum Tweet length for which a dataset of 5000 characters was available.

An example tweet from the corpus is shown below in Table 4-1.

Table 4-1 Example Tweet Anatomy

1. Sentiment	2. Tweet ID	3. Datetime	4. Twitter Specific	5. Author	6. Text Body
0	1467811592	Mon Apr 06 22:20:03	NO_QUERY	mybirch	Need a hug

Fields 2, 3, 4 and 5 of the datasets were discarded retaining only field 6, the text of the tweet, and field 1, which contained the sentiment assigned to the text by the Sentiment140 project, negative sentiment encoded as a 0, positive sentiment encoded as a 4. The coding for positive sentiment was changed to a 1 for all positive tweets in the data, as some of the used classifier implementations expect input binary class information to be encoded as either 0 or 1.

The words for each tweet in the length specific subsets were counted and averaged. These results are presented in Table 4-2 below. Interestingly, mean word length decreases gradually as text length increases.

Table 4-2 Mean and Standard Deviation of Word Counts by Text Length

Text Length	Mean Word Length	StdDev of Word Length
10	7.38	2.6
15	6.34	2.7
20	6.29	1.9
25	6.36	2.0
30	6.17	1.8
35	6.07	1.5
40	6.04	1.4
45	5.97	1.3
50	5.95	1.3
80	5.78	1.1
110	5.68	0.9
138	5.61	2.1
Overall	**6.14**	**1.9**

4.2. Data Preparation

The length specific data sets, containing the original tweet message bodies were enhanced by the methods outlined below, each enhancement adding a new attribute by concatenating the enriched features with the 'Lemmatised' data set as described below.

After enrichment, the text features available for analysis were:

- Original – the original text of the tweet as extracted from the Sentiment140 dataset.
- Cleaned – the original text having punctuation and stop words removed, and twitter specific strings (e.g. hashtags, URLs) replaced with standard tokens. Cleaning is intended to enhance through the removal of features which do not contribute to classification accuracy and words that are either so common or so infrequent that they are useless as class discriminators.
- Lemmatised – the cleaned set (above) lemmatised using the NLTK python library. Lemmatization is intended to enhance by increasing the statistical similarity of words having the same root, through substitution of words with their root-words (lemmas).
- Bigrams – enhanced by appending all bigrams from the lemmatised tweet back to the lemmatized tweet. Bigrams provide a degree of context to a classifier – frequent pair

occurrences become features in their own right so increasing the specificity of surface matching.

- Synonyms – enhanced by appending all available wordnet synonyms for each word in the lemmatised tweet to the lemmatized tweet. The addition of synonym is intended to act as a rudimentary method of 'concept mining' by increasing the chances of a match between tweets sharing synonyms (common concepts), but without common words.
- Hypernyms – enhanced by appending all available wordnet hypernyms for each word in the lemmatised tweet back to the lemmatized tweet. Intended to act in a similar way to synonyms but with the ability to match tweets with greater concept specificity.
- Hyponyms – enhanced by appending all available wordnet hyponyms for each word in the lemmatised tweet back to the lemmatized tweet. Intended to act in a similar way to synonyms but with the ability to match tweets with greater concept generality.
- Wiki Words – enhanced by appending all available words in all the 'labels' contained in the top five Wikipedia hits for each word in the lemmatised text back to the lemmatised text. Designed to be an explicit match of common concepts (taxonomic classes and categories) which occur in the returned Wikipedia metadata.
- Wiki Phrases – enhanced by appending all available 'labels', each treated as an indivisible string (n-gram), from the top five Wikipedia hits for each word in the lemmatised text back to the lemmatised text. Designed to be an explicit match of common concepts (taxonomic classes and categories) which occur in the returned Wikipedia metadata, but with greater contextual power during the classification stage due to the specificity of the n-gram rather than the previous, more general, bag of words approach.
- Wiki Bigrams - enhanced by appending all available 'labels', each treated as an indivisible string (n-gram), from the top five Wikipedia hits for each bigram in the lemmatised text back to the lemmatised text. Designed to be an explicit match of common concepts (taxonomic classes and categories) which occur in the returned Wikipedia metadata, but with greater contextual power during both the classification stage due to the specificity of the n-gram, and the enhancement phase, due to the specificity of the input bigram, than either of the previous, more general approaches.

An illustrative example of the results of all of the various enhancements of one single tweet, "**@projectkpaz sounds good**", is given in appendix 8.3.

4.3. Trial Modelling

Prior to full modelling and classification execution, some pilot work was undertaken to check the basic health and functionality of the code and data. This resulted in two major changes. The first was a change to the default setting for the LSA classifier's retained vectors parameter: the value of 500 retained vectors in the LSA dimensionality reduction step was chosen after some preliminary experimentation which suggested that retaining any more than 500 vectors yielded diminishing returns in terms of classifier performance, while retaining fewer than 500 vectors resulted in a noticeable fall-off in performance. This judgement was made on visual inspection of the plot in Figure 4-2 below, rather than on any formal statistical analysis. This is in line with the stated strategy in section 1.4 above: the purpose of this project is not to formally optimise the classifiers.

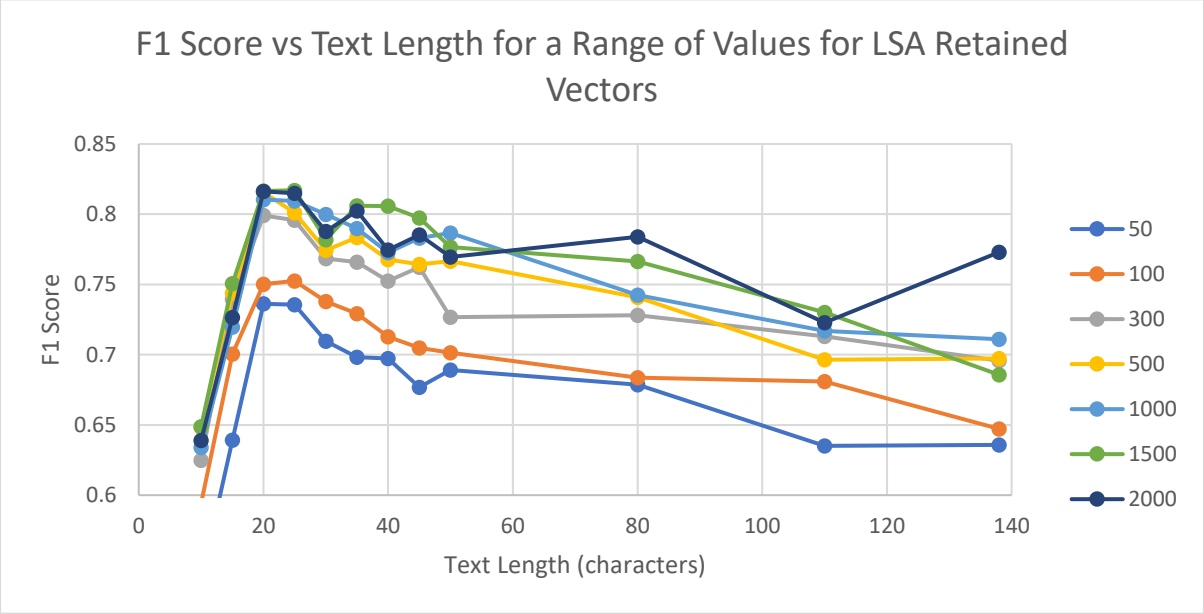

Figure 4-2 Relationship between LSA Retained Vectors and Classifier Performance

The second change was to augment the initial group of 5 length specific datasets (138, 110, 80, 50, 20 characters) after trial plots indicated that there were rapid changes in the accuracy of classification between the 50-character set and the 20-character set. As this is exactly the response that the experiment was designed to capture, further data sets at character length intervals of five characters were created and added to the experiment, as were sets of 15 characters and <= 10 characters, in order to improve resolution in the area of interest. The final length categories were 138, 110, 80, 50, 45, 40, 35, 30, 25, 20, 15 and <=10 characters.

4.4. Modelling and Classification

Following adjustments informed by trial modelling each of the ten types of processed tweet was classified by the following methods:

- Naïve-Bayes using the built-in routines from the scikit-learn python library.
- Support Vector Machine (SVM) using the built-in routines from the scikit-learn python library.
- Latent semantic analysis, keeping only the top 500 vectors during the dimensionality reduction step, using the built-in routines from the scikit-learn python library followed by Support Vector Machine (SVM) using the built-in routines from the scikit-learn python library.

Each classification was repeated 100 times, with the data set being randomly shuffled between repeats. Each classification run used Monte Carlo cross-validation, randomly splitting the set 90% into training data used to train the supervised learning model and 10% into test data used to assess the classification performance of that model. After each classification run, model performance was measured using the F1 Score from the classification of the test data and the score recorded. The F1 Score is calculated as the weighted average of precision and recall measures. The F1 Score results for each combination of classifier, enhancement method and tweet length were averaged (mean of 100 runs).

The averaged F1 Scores for each combination of classifier, enhancement method and tweet length are presented below in tabular and graphic format.

4.5. Results for Naïve-Bayes Classification

The mean F1 Scores, taken from 100 separate modelling runs, for each combination of enhancement and text length are shown in Table 4-3 below.

Table 4-3 Classification Performance (F1 Score) Results - Naïve Bayes

F1 Score	Enhancement										
Text Length	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words	Mean
10	0.795	0.801	0.769	0.776	0.799	0.861	0.796	0.789	0.795	0.775	0.796
15	0.864	0.854	0.804	0.804	0.853	0.894	0.830	0.826	0.829	0.798	0.836
20	0.900	0.892	0.825	0.825	0.891	0.913	0.846	0.869	0.845	0.815	0.862
25	0.904	0.896	0.819	0.808	0.894	0.906	0.826	0.881	0.842	0.814	0.859
30	0.918	0.889	0.810	0.796	0.886	0.904	0.827	0.871	0.826	0.800	0.853
35	0.924	0.892	0.793	0.777	0.891	0.905	0.813	0.874	0.814	0.791	0.848
40	0.936	0.895	0.782	0.765	0.893	0.908	0.805	0.874	0.806	0.786	0.845
45	0.938	0.895	0.772	0.758	0.894	0.906	0.797	0.871	0.805	0.778	0.841
50	0.941	0.893	0.766	0.739	0.891	0.908	0.796	0.860	0.799	0.771	0.836
80	0.959	0.905	0.756	0.713	0.903	0.915	0.786	0.862	0.802	0.761	0.836
110	0.960	0.906	0.756	0.712	0.904	0.918	0.785	0.849	0.798	0.770	0.836
138	0.960	0.912	0.769	0.704	0.910	0.921	0.788	0.844	0.799	0.753	0.836
Total	11.00	10.63	9.42	9.18	10.61	10.86	9.70	10.27	9.76	9.41	10.08

Casual inspection does not reveal an immediately apparent trend. The data in columns represents the performance of the classifier on each particular enhancement, and the 'Total' number at the bottom of each column is the sum of all F1-Scores for that column.

That sum of all F1 Scores for a given enhancement may be thought of as a crude measure of the area-under-the-curve for the performance of that enhancement, and the figure at bottom right is the mean of those areas, and as such may be taken as a very crude measure of overall performance of the particular classifier. In both cases a greater area implies better performance.

These F1 Score data are represented graphically in Figure 4-3 below, where some trends are becoming apparent.

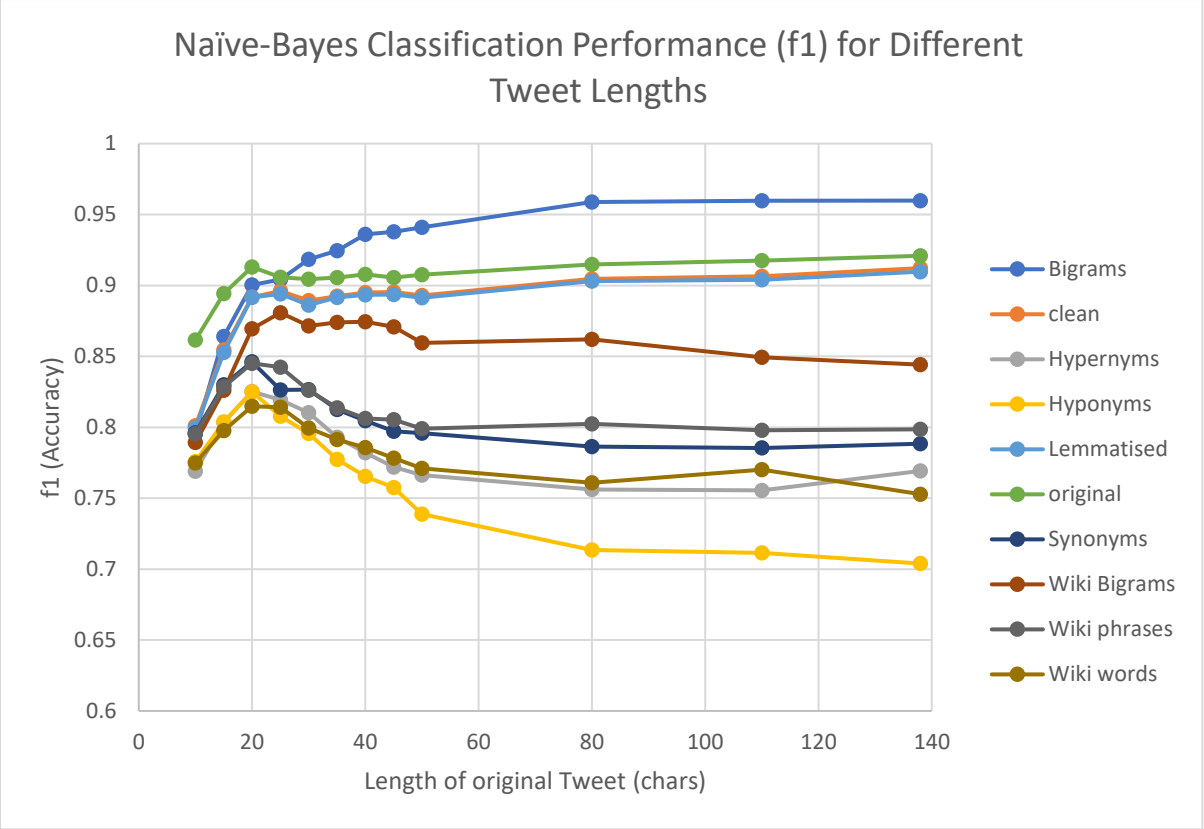

Figure 4-3 Classification Performance (F1 Score) Results - Naïve Bayes

The graphical plot in Figure 4-3 above illustrates some notable features.

There appears to be an area of relatively stable F1 score performance with respect to text lengths between 80 and 138 characters, for all enhancements.

All enhancements exhibit a steep fall-off in performance for text lengths below 20 characters.

Recalling the distinction made in section 3.2 between additive and non-additive enhancements, it may be noted that all the non-additive enhancements (original, lemmatized, cleaned and bigrams) show a weak tendency for performance to increase with text length, while the additive enhancements seems to show a fall in performance as length increases from 20 to 80 characters.

Table 4-4, below, shows the standard deviations associated with the 100 F1 scores for each enhancement-length combination. In very general terms, variance appears to be relatively low.

Table 4-4 Standard Deviations of F1 Score Results - Naïve Bayes

StdDev of F1	Enhancement									
Text Length	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words
10	0.0031	0.0033	0.0034	0.0063	0.0036	0.0034	0.0035	0.0032	0.0036	0.0035
15	0.0025	0.0025	0.0028	0.0029	0.0025	0.0026	0.0028	0.0038	0.0029	0.0031
20	0.0055	0.0024	0.0029	0.0033	0.0024	0.0025	0.0029	0.0043	0.0031	0.0032
25	0.0022	0.0025	0.0028	0.0037	0.0025	0.0035	0.0026	0.0029	0.0024	0.0035
30	0.0033	0.0028	0.0033	0.0038	0.0027	0.0027	0.0027	0.0029	0.0029	0.0035
35	0.0022	0.0027	0.0033	0.0032	0.0029	0.0030	0.0033	0.0031	0.0031	0.0027
40	0.0028	0.0026	0.0036	0.0035	0.0027	0.0026	0.0030	0.0031	0.0033	0.0035
45	0.0023	0.0026	0.0037	0.0033	0.0027	0.0025	0.0035	0.0029	0.0033	0.0032
50	0.0024	0.0027	0.0043	0.0038	0.0028	0.0032	0.0037	0.0036	0.0032	0.0035
80	0.0023	0.0031	0.0039	0.0035	0.0026	0.0030	0.0032	0.0032	0.0031	0.0044
110	0.0023	0.0029	0.0038	0.0041	0.0029	0.0033	0.0033	0.0037	0.0035	0.0037
138	0.0022	0.0027	0.0039	0.0042	0.0028	0.0029	0.0040	0.0035	0.0038	0.0048

Using data from tables Table 4-3 and Table 4-4 the standard deviation can be expressed as a percentage of the underlying mean F1 score for each combination of enhancement and text length, and Figure 4-4 shows a plot of this measure across enhancements and text lengths. It can be seen that, for all points, the standard deviations are less than 1% of the underlying mean value.

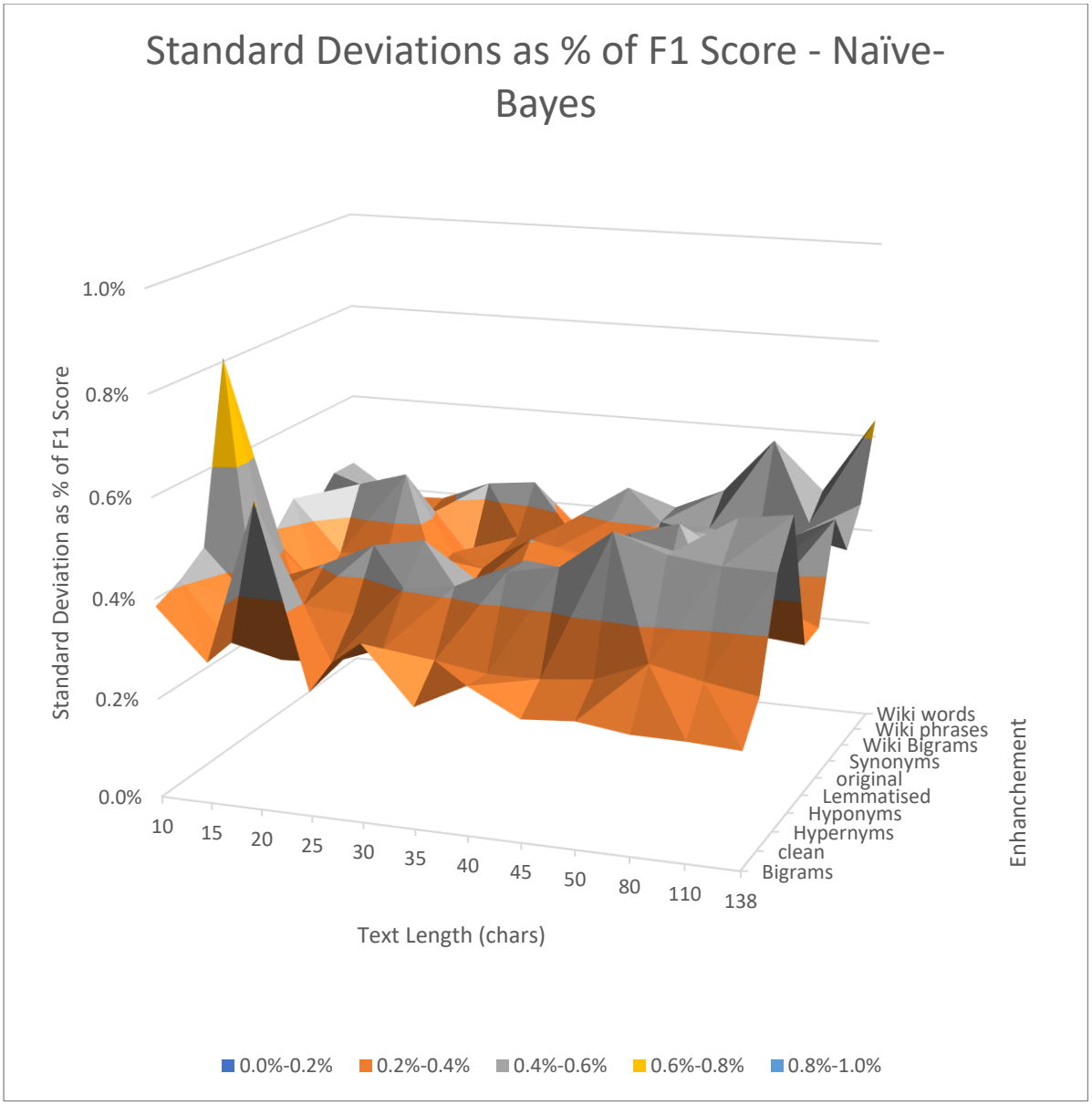

Figure 4-4 Standard Deviations as % of F1 Score - Naïve-Bayes

Figure 4-5, below, replots this same data on a rescaled vertical axis in order to facilitate a visual comparison between this data and the data from the SVM and LSA classifiers which are presented in figures Figure 4-7 and Figure 4-9, respectively, later in this section.

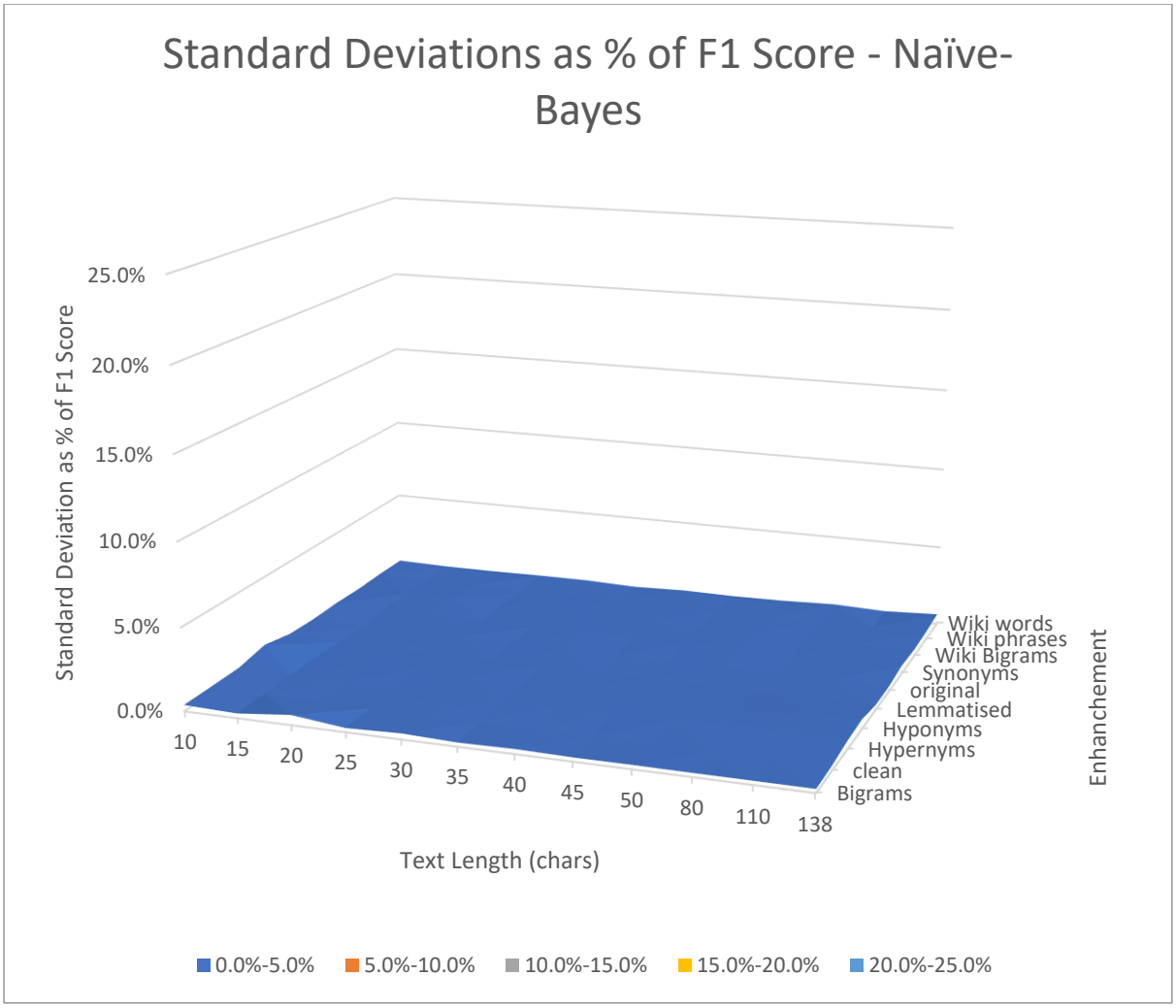

Figure 4-5 Standard Deviations as % of F1 Score - Naïve-Bayes - Expanded Scale

4.6. Results for Support Vector Machine Classification

The mean F1 Scores, taken from 100 separate modelling runs, for each combination of enhancement and text length are shown below.

Table 4-5 Classification Performance (F1 Score) Results – Support Vector Machine

F1 Score	Enhancement										
Text Length	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words	Mean
10	0.677	0.796	0.762	0.716	0.794	0.800	0.776	0.666	0.734	0.715	0.744
15	0.810	0.851	0.835	0.817	0.848	0.883	0.843	0.786	0.816	0.777	0.827
20	0.871	0.899	0.857	0.838	0.898	0.908	0.871	0.849	0.833	0.786	0.861
25	0.884	0.899	0.844	0.819	0.898	0.912	0.861	0.855	0.825	0.773	0.857
30	0.893	0.899	0.831	0.800	0.895	0.911	0.853	0.848	0.802	0.751	0.848
35	0.906	0.900	0.823	0.793	0.897	0.915	0.845	0.852	0.797	0.744	0.847
40	0.922	0.902	0.814	0.778	0.899	0.917	0.840	0.854	0.779	0.715	0.842
45	0.925	0.900	0.809	0.772	0.899	0.918	0.835	0.857	0.783	0.716	0.841
50	0.930	0.898	0.799	0.754	0.897	0.917	0.825	0.848	0.774	0.704	0.835
80	0.958	0.902	0.781	0.741	0.898	0.919	0.816	0.847	0.742	0.688	0.829
110	0.960	0.902	0.768	0.728	0.895	0.920	0.804	0.829	0.701	0.640	0.815
138	0.959	0.897	0.765	0.712	0.892	0.916	0.798	0.830	0.688	0.628	0.808
Total	10.696	10.645	9.690	9.266	10.611	10.836	9.967	9.920	9.274	8.636	9.954

Again, casual inspection does not reveal an immediately apparent trend. The data in columns represents the performance of the classifier on each particular enhancement, and the 'Total' number at the bottom of each column is the sum of all F1-Scores for that column.

As mentioned above, that sum of all F1 Scores for a given enhancement may be thought of as a crude measure of the area-under-the-curve for the performance of that enhancement, and the figure at bottom right is the mean of those areas, and as such may be taken as a very crude measure of overall performance of the particular classifier. In both cases a greater area implies better performance.

These F1 Score data are represented graphically in Figure 4-6 below, where some trends are becoming apparent.

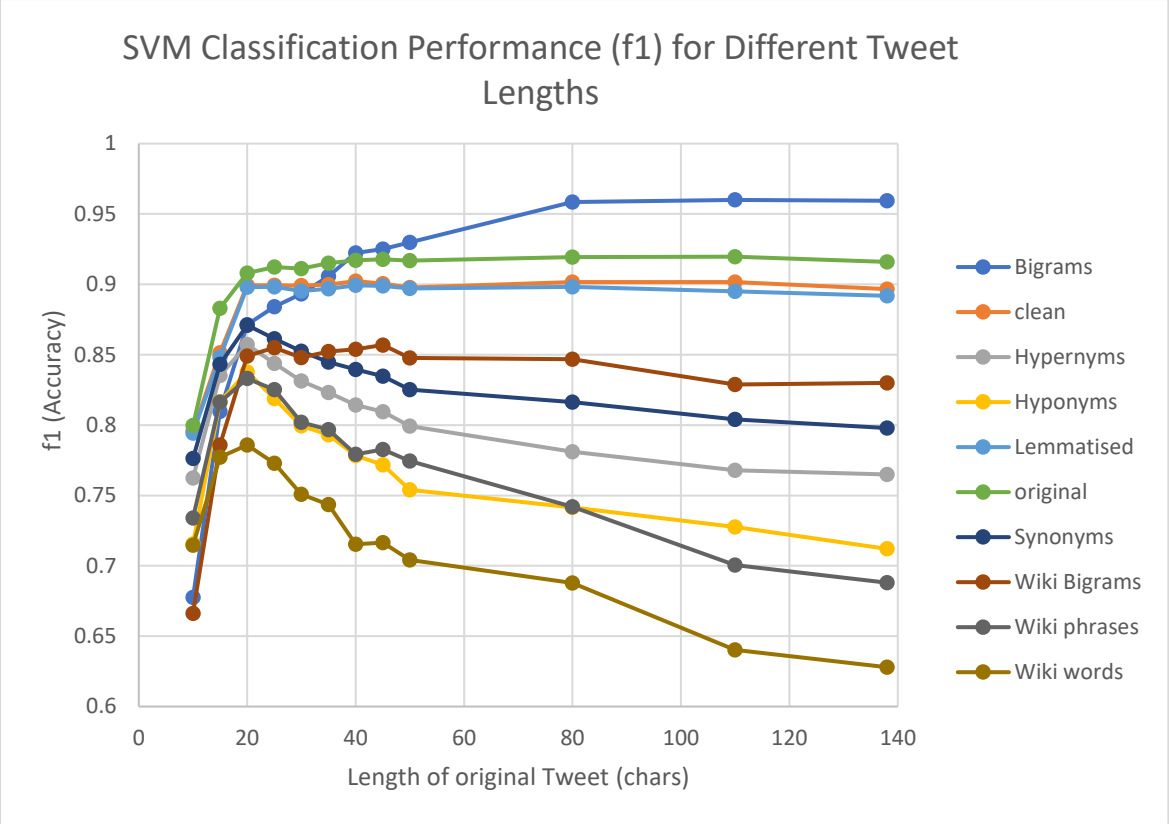

Figure 4-6 Classification Performance (F1 Score) Results - Support Vector Machine

The graphical plot in Figure 4-6 above bears a strong similarity to Figure 4-3, the corresponding plot for naïve-Bayes, and illustrates some notable features.

Recalling the distinction made in section 3.2 between additive and non-additive enhancements, it may be noted that the additive enhancements seems to show a fall in performance as length increases from 20 to 138 characters.

There appears to be an area of relatively stable F1 score performance with respect to text lengths between 80 and 138 characters, for non-additive enhancements.

All enhancements exhibit a steep fall-off in performance for text lengths below 20 characters.

Table 4-6, below, shows the standard deviations associated with the 100 F1 scores for each enhancement-length combination. In very general terms variances appear to be higher than the corresponding naïve-Bayes variances.

Table 4-6 Standard Deviations of F1 Score Results - Support Vector Machine

StdDev of F1	Enhancement									
Text Length	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words
10	0.0055	0.0060	0.0288	0.0252	0.0061	0.0093	0.0292	0.0074	0.0397	0.0350
15	0.0101	0.0075	0.0097	0.0127	0.0092	0.0169	0.0093	0.0108	0.0234	0.0332
20	0.0045	0.0052	0.0063	0.0094	0.0079	0.0039	0.0063	0.0046	0.0242	0.0344
25	0.0054	0.0072	0.0071	0.0113	0.0078	0.0039	0.0070	0.0067	0.0294	0.0453
30	0.0060	0.0052	0.0108	0.0111	0.0095	0.0034	0.0074	0.0107	0.0388	0.0371
35	0.0074	0.0080	0.0097	0.0122	0.0111	0.0040	0.0079	0.0098	0.0398	0.0426
40	0.0064	0.0071	0.0117	0.0139	0.0174	0.0041	0.0101	0.0112	0.0657	0.0759
45	0.0059	0.0094	0.0133	0.0158	0.0095	0.0033	0.0102	0.0144	0.0514	0.0679
50	0.0057	0.0102	0.0143	0.0159	0.0069	0.0035	0.0120	0.0148	0.0504	0.0713
80	0.0034	0.0098	0.0204	0.0179	0.0105	0.0051	0.0178	0.0184	0.0909	0.0805
110	0.0026	0.0082	0.0288	0.0218	0.0092	0.0046	0.0164	0.0275	0.1339	0.1171
138	0.0026	0.0115	0.0287	0.0278	0.0133	0.0059	0.0205	0.0226	0.1452	0.1200

Using data from tables Table 4-5 and Table 4-6 the standard deviation can be expressed as a percentage of the underlying mean F1 score for each combination of enhancement and text length, and Figure 4-7 shows a plot of this measure across enhancements and text lengths. It can be seen that this plot shows some striking differences in comparison with Figure 4-5, the corresponding naïve-Bayes plot.

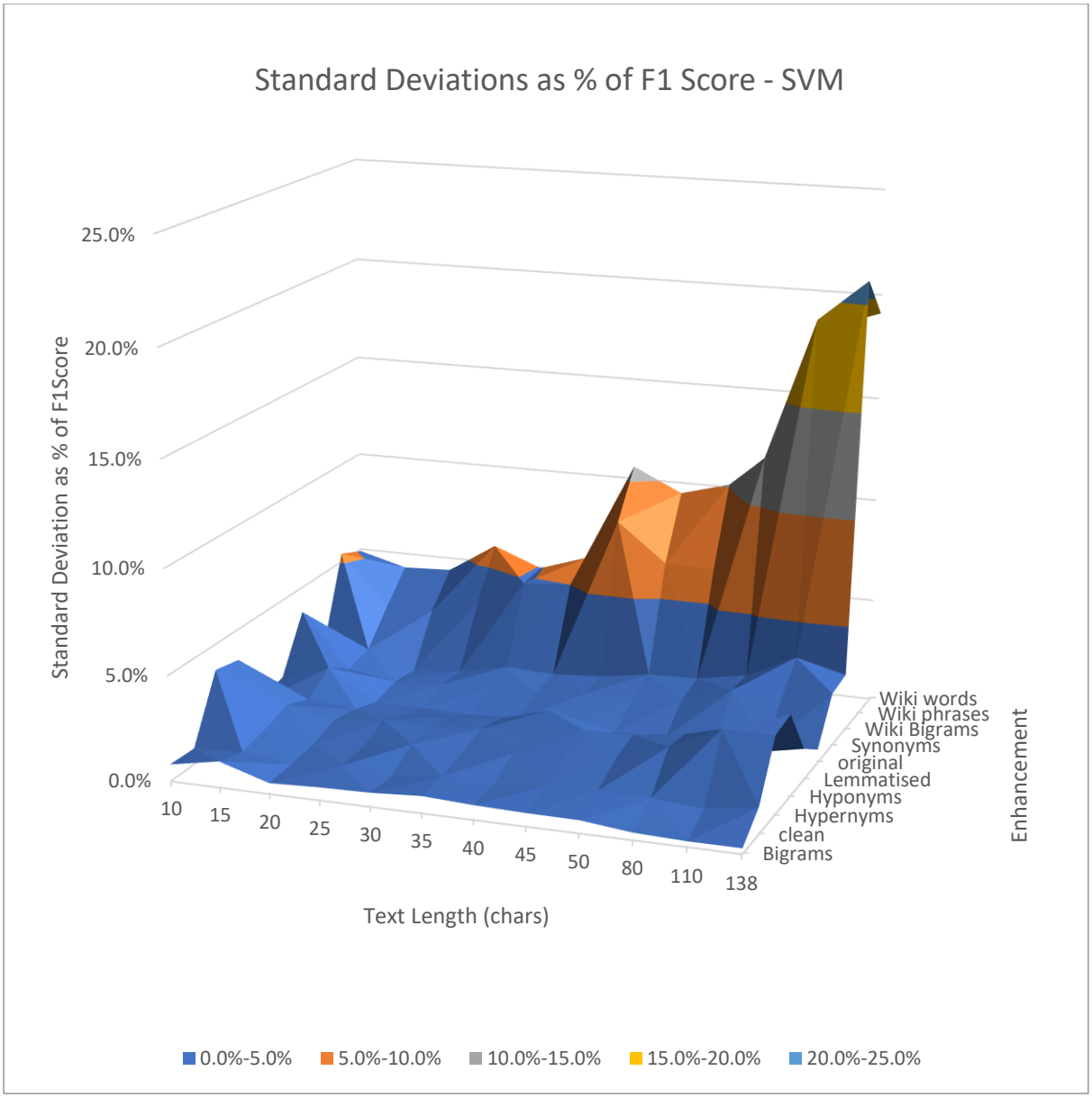

Figure 4-7 Standard Deviations as % of F1 Score - SVM

Overall, standard deviations, as a percentage of underlying mean, are higher than for naïve-Bayes. There is an upward trend as text length increases, in general, but the most salient feature is the very high variance for the Wiki Words and Wiki Phrases enhancements where the standard deviations typically exceed 5% of underlying mean, and peak at >20% at longer text lengths.

4.7. Results for Latent Semantic Analysis / SVM Classification

The mean F1 Scores, taken from 100 separate modelling runs, for each combination of enhancement and text length are shown below.

Table 4-7 Classification Performance (F1 Score) Results – Latent Semantic Analysis

F1 Score	Enhancement										
Text Length	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words	Mean
10	0.647	0.644	0.697	0.648	0.644	0.651	0.693	0.645	0.665	0.662	0.660
15	0.743	0.749	0.749	0.720	0.751	0.733	0.766	0.745	0.740	0.728	0.742
20	0.808	0.805	0.804	0.780	0.811	0.808	0.807	0.804	0.797	0.787	0.801
25	0.801	0.799	0.787	0.770	0.798	0.806	0.801	0.796	0.791	0.782	0.793
30	0.783	0.779	0.764	0.746	0.782	0.784	0.787	0.775	0.770	0.755	0.773
35	0.775	0.775	0.756	0.737	0.773	0.776	0.770	0.766	0.755	0.754	0.764
40	0.772	0.764	0.745	0.724	0.768	0.774	0.770	0.756	0.760	0.750	0.758
45	0.775	0.769	0.733	0.721	0.771	0.768	0.761	0.761	0.746	0.737	0.754
50	0.762	0.759	0.726	0.694	0.761	0.763	0.753	0.753	0.738	0.731	0.744
80	0.753	0.747	0.693	0.667	0.746	0.752	0.728	0.729	0.714	0.713	0.724
110	0.718	0.719	0.689	0.650	0.714	0.721	0.710	0.695	0.692	0.694	0.700
138	0.709	0.708	0.685	0.624	0.704	0.703	0.702	0.688	0.678	0.671	0.687
Total	9.047	9.017	8.828	8.479	9.024	9.038	9.048	8.914	8.845	8.764	8.900

Once again, casual inspection does not reveal an immediately apparent trend. The data in columns represents the performance of the classifier on each particular enhancement, and the 'Total' number at the bottom of each column is the sum of all F1-Scores for that column.

Again, that sum of all F1 Scores for a given enhancement may be thought of as a crude measure of the area-under-the-curve for the performance of that enhancement, and the figure at bottom right is the mean of those areas, and as such may be taken as a very crude measure of overall performance of the particular classifier. In both cases a greater area implies better performance.

These F1 Score data are represented graphically in Figure 4-8 below, where some trends are becoming apparent.

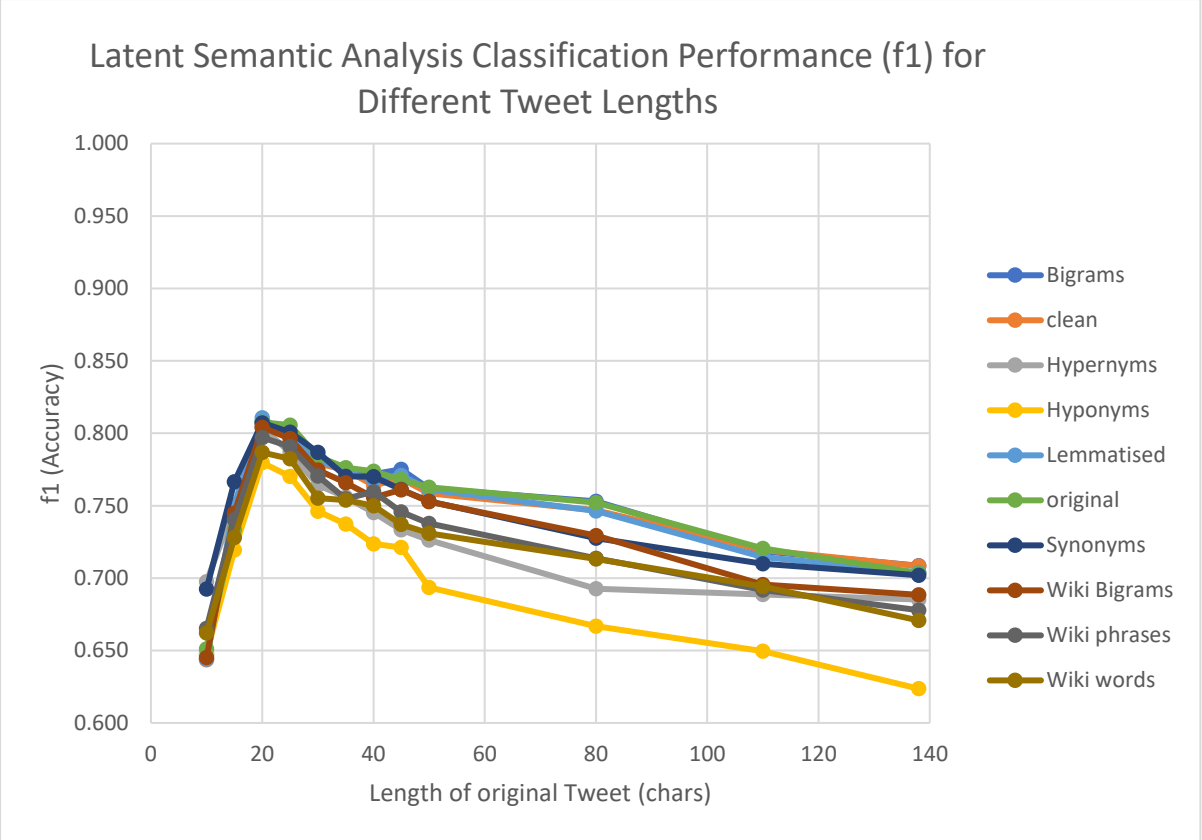

Figure 4-8 Classification Performance (F1 Score) Results - Latent Semantic Analysis

The graphical plot in Figure 4-8 above bears only some similarity to Figure 4-3 and Figure 4-6, the corresponding plots for naïve-Bayes and SVM, and illustrates some notable features.

Recalling the distinction made in section 3.2 between additive and non-additive enhancements, it may now be noted that no distinct differences are apparent as performance decreases as length increases from 20 to 138 characters.

There appears to be no area of relatively stable F1 score performance with respect to text lengths between 80 and 138 characters, in contrast with the naïve-Bayes and SVM classifiers.

Again, all enhancements exhibit a steep fall-off in performance for text lengths below 20 characters.

In general terms, F1 scores are lower than for naïve-Bayes or SVM classifiers.

Table 4-8, below, shows the standard deviations associated with the 100 F1 scores for each enhancement-length combination. In very general terms, variances appear to be higher than the corresponding naïve-Bayes variances, but to have lower peak values that the support vector models.

Table 4-8 Standard Deviations of F1 Score Results - Latent Semantic Analysis

StdDev of F1	Enhancement									
Text Length	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words
10	0.0215	0.0247	0.0214	0.0252	0.0213	0.0213	0.0210	0.0236	0.0137	0.0166
15	0.0311	0.0291	0.0281	0.0262	0.0246	0.0320	0.0302	0.0209	0.0182	0.0260
20	0.0119	0.0137	0.0098	0.0127	0.0121	0.0144	0.0132	0.0140	0.0106	0.0123
25	0.0119	0.0172	0.0131	0.0123	0.0143	0.0111	0.0109	0.0132	0.0101	0.0107
30	0.0176	0.0188	0.0193	0.0185	0.0196	0.0185	0.0131	0.0158	0.0145	0.0179
35	0.0191	0.0215	0.0191	0.0150	0.0228	0.0228	0.0140	0.0221	0.0188	0.0154
40	0.0171	0.0266	0.0223	0.0225	0.0216	0.0189	0.0140	0.0258	0.0189	0.0200
45	0.0172	0.0237	0.0215	0.0158	0.0220	0.0254	0.0166	0.0198	0.0224	0.0255
50	0.0276	0.0283	0.0288	0.0223	0.0285	0.0262	0.0233	0.0245	0.0261	0.0266
80	0.0332	0.0305	0.0477	0.0301	0.0305	0.0260	0.0265	0.0373	0.0343	0.0240
110	0.0318	0.0283	0.0478	0.0398	0.0366	0.0332	0.0402	0.0391	0.0395	0.0421
138	0.0412	0.0343	0.0377	0.0549	0.0362	0.0460	0.0326	0.0482	0.0453	0.0499

Using data from Table 4-7 and Table 4-8 the standard deviation can be expressed as a percentage of the underlying mean F1 score for each combination of enhancement and text length, and Figure 4-9 shows a plot of this measure across enhancements and text lengths. It can be seen that this plot shows some striking differences to Figure 4-5, the corresponding naïve-Bayes plot, and to Figure 4-7, the corresponding SVM plot.

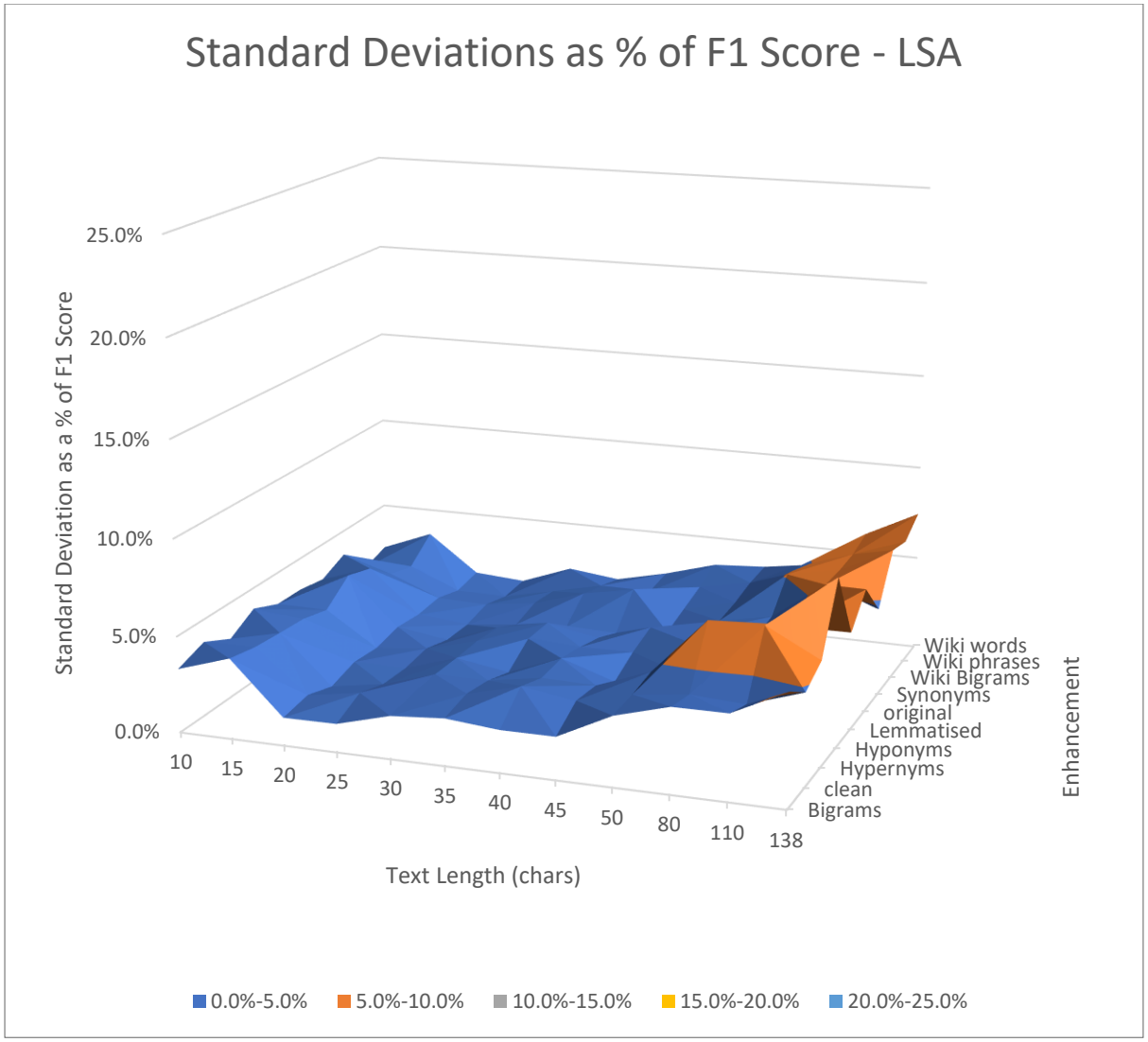

Figure 4-9 Standard Deviations as % of F1 Score - LSA

Overall, standard deviations, as a percentage of underlying mean, are higher than for naïve-Bayes. There is a noticeable overall upward trend as text length increases, in general, but no marked trend between enhancements as was seen in the corresponding SVM plot.

4.8. Comparisons of Enhancements and Classifiers

Although performance comparisons between enhancements and between classifiers is not a primary goal of this work, some informal conclusions may be inferred from the data already collected.

Gathering together the 'Total' rows from Table 4-3, Table 4-5 and Table 4-7, a summary table of the crude 'area-under-the-curve' data is presented below.

Table 4-9 Areas under the Performance Curve for Enhancements and Classifiers

	Enhancement									
Classifier	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words
NB	11.000	10.632	9.421	9.176	10.610	10.859	9.695	10.271	9.761	9.411
SVM	10.696	10.645	9.690	9.266	10.611	10.836	9.967	9.920	9.274	8.636
LSA	9.047	9.017	8.828	8.479	9.024	9.038	9.048	8.914	8.845	8.764

By sorting and plotting the data from Table 4-9, Figure 4-10 can be created.

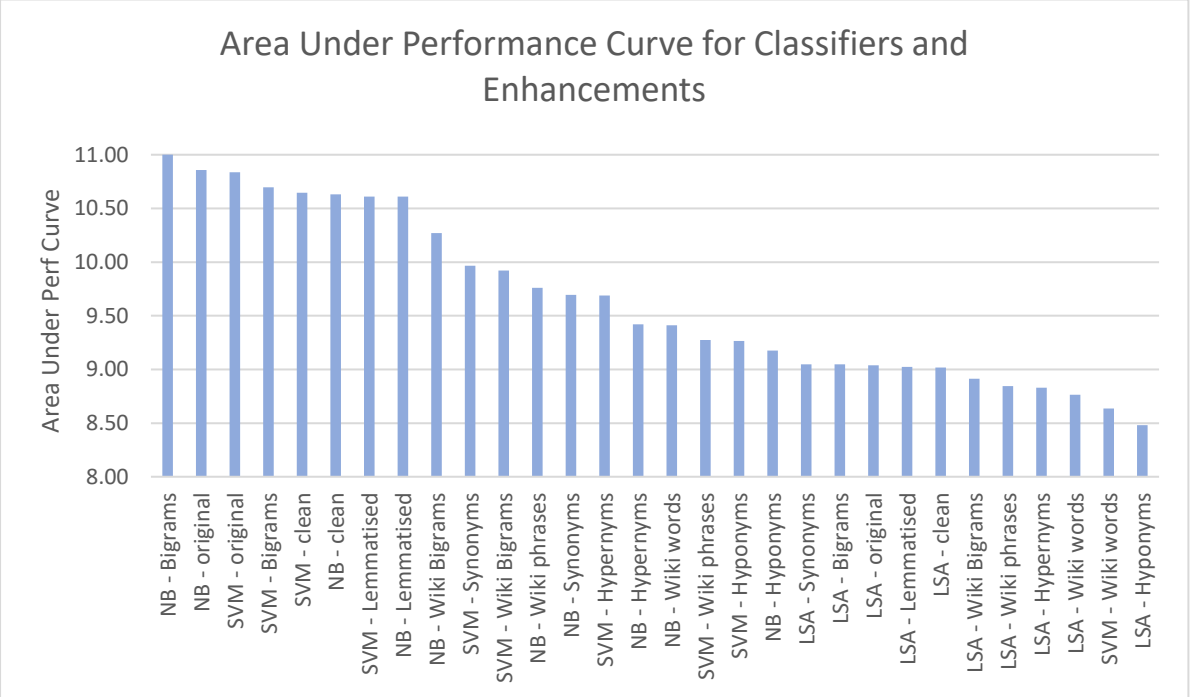

Figure 4-10 Areas under the Performance Curve for Enhancements and Classifiers

Figure 4-10 presents the dimensionless (arbitrary) 'area under the performance curve' measures for all classifier-enhancement combinations, sorted in descending order of performance from left to right.

Immediately apparent is the relatively poor performance of the LSA classifier.

Recalling, once again, the distinction made in section 3.2 between additive and non-additive enhancements, it may be noted that, of the top eight combinations, all use non-additive enhancements and that, discounting LSA results, no additive enhancement out-performs any non-additive enhancement, suggesting that the additive enhancements (Wordnet, Wikipedia) are under-performing the simpler, non-additive, enhancements.

At the risk over over-simplifying a complex issue with a single unsophisticated measure, but in the interest of completeness, Table 4-10, below, presents the sum of the areas under all the enhancement performance curves for each classifier. This may be tentatively interpreted as a measure of overall F1 score classification performance across all text lengths and all enhancements for each classifier.

Table 4-10 Total Area Under All Curves by Classifier

Classifier	Total Area under all curves
Naïve-Bayes	100.8
SVM	99.5
LSA	89.0

It can be seen that naïve-Bayes seems to very slightly out-perform the support vector classifier, and the LSA method seems to lag the others in terms of performance. Given the discussion in section 2.1.1, this performance for naïve-Bayes should not be considered particularly surprising.

4.9. Summary

The average performance of three classification methods was tested on short texts of twelve different lengths, having 10 different enhancements applied to the texts before classification. Tabular and graphical results of the F1 Score of these classification operations were presented along with results on the variance of the various methods and their relative overall performance.

Chapter 5 will analyse these results from a statistical perspective.

5. EVALUATION / ANALYSIS

This chapter will discuss statistical testing of the results presented in Chapter 4 and draw quantitative conclusions on the specific hypotheses set out in Section 1.3 above.

All required code for analysis of the results data was written in R 3.3.3 (Another Canoe - release date 06/03/2017). Required R packages will be detailed below. All statistical tests associated with the project were conducted at a confidence level of 95% unless otherwise stated.

5.1. Homogeneity of Variance

The three result sets, one for each classifier, were first tested for homogeneity of variance using Levene's test as implemented by the General Linear Model in R. A p-value of less than 0.05 indicates that there is insufficient homogeneity of variance (at a 95% confidence) to ensure validity from the standard form of ANOVA testing.

Figure 5-1 Levene's Test for Homogeneity of Variance on Naïve Bayes F1-Score

```
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group   119  7.2771 < 2.2e-16 ***
      11880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5-2 Levene's Test for Homogeneity of Variance on SVM F1-Score

```
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group   119  30.047 < 2.2e-16 ***
      11880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5-3 Levene's Test for Homogeneity of Variance on LSA F1-Score

```
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group   119  11.592 < 2.2e-16 ***
      11880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In all three cases, Levene's test indicated that the heterogeneity of variance was sufficient to require more robust analysis that that available from standard ANOVA. The robust method chosen was a trimmed-means procedure described by Wilcox (Wilcox & Keselman, 2003) and implemented in the R package WRS2.

5.2. Robust 2-Way ANOVA

Results of Wilcox's robust trimmed means 2-way independent ANOVA for all three classifiers are shown below.

Table 5-1 Results of Robust 2-way ANOVA testing on Naïve Bayes Classification

	Test Value	p-Value
Enhancement	2642891.7	**0.001**
Length	200056.2	**0.001**
Interaction	518945.2	**0.001**
		p-values < 0.05 are significant

The p-values for the 2-way ANOVA for naïve-Bayes classification indicate that significant differences in F1 Score occur both between text lengths and between enhancement methods. The interaction between text length and enhancement is also significant.

These results indicate that, (at 95% confidence):

1) taken across all lengths, performance is significantly influenced by enhancement. Neither hypothesis was contingent on this relationship.
2) taken across all enhancements, performance is significantly influenced by text length. On its own, this finding cannot be used to reject hypothesis1 because enhancements have not been individually tested.
3) the change of performance with respect to length is statistically different between enhancements. This finding is sufficient to reject hypothesis 2 for the naïve-Bayes classifier.

Table 5-2 Results of Robust 2-way ANOVA testing on SVM Classification

	Test Value	p-Value
Enhancement	119850.24	**0.001**
Length	25036.01	**0.001**
Interaction	121528.29	**0.001**
		p-values < 0.05 are significant

The p-values for the 2-way ANOVA for Support Vector Machine classification indicate that significant differences in F1 Score occur both between text lengths and between enhancement methods. The interaction between text length and enhancement is also significant.

These results indicate that, (at 95% confidence):

1) taken across all lengths, performance is significantly influenced by enhancement. Neither hypothesis was contingent on this relationship.
2) taken across all enhancements, performance is significantly influenced by text length. On its own, this finding cannot be used to reject hypothesis1 because enhancements have not been individually tested.
3) the change of performance with respect to length is statistically different between enhancements. This finding is sufficient to reject hypothesis 2 for the Support Vector Machine classifier.

Table 5-3 Results of Robust 2-way ANOVA testing on LSA Classification

	Test Value	**p-Value**
Enhancement	4254.151	**0.001**
Length	48572.718	**0.001**
Interaction	2413.403	**0.001**
p-values < 0.05 are significant		

The p-values for the 2-way ANOVA for Latent Semantic Analysis / SVM classification indicate that significant differences in F1 Score occur both between text lengths and between enhancement methods. The interaction between text length and enhancement is also significant.

These results indicate that, (at 95% confidence):

1) taken across all lengths, performance is significantly influenced by enhancement. Neither hypothesis was contingent on this relationship.
2) taken across all enhancements, performance is significantly influenced by text length. On its own, this finding cannot be used to reject hypothesis1 because enhancements have not been individually tested.

3) the change of performance with respect to length is statistically different between enhancements. This finding is sufficient to reject hypothesis 2 for the Latent Semantic Analysis classifier.

5.3. Robust 1-Way ANOVA and Trend Testing

Following 2-Way ANOVA, the individual result sets for combinations of classifier and enhancement were tested using Wilcox's robust 1-Way ANOVA (based on Levene's test results showing an insufficient homogeneity of variance) and Jonckheere-Terpstra's 2-tailed test for the presence of a trend using text length as the independent variable (Jonckheere, 1954).

In all combinations of classifier and enhancement significant differences in F1 Score with changes in text-length were detected by the 1-Way ANOVA.

In all combinations of classifier and enhancement significant trends in F1 Score with changing text-length were detected by Jonckheere-Terpstra's 2-tailed test.

Both of these results indicate that for all classifiers and for all enhancements the text length has a significant effect, at the 95% confidence level, on the classification performance, and these individual results indicate that, as hinted at by the 2-way ANOVA test, hypothesis 1 may be rejected for all enhancement – classifier combinations.

Tabular data for the 1-way analyses are presented in Table 5-4 below. (All p-values less than 0.05 are considered to be significant at the 95% confidence level.)

Table 5-4 Results of Individual 1-way ANOVA and Jonckheere-Terpstra Tests

		p-value		
Classifier	**Enhancement**	**Levene**	**Robust 1-way Anova**	**Jonckheere-Terpstra**
NB	Lemmatised	1.027E-03	0.0000	< 2.2e-16
NB	Synonyms	8.927E-04	0.0000	< 2.2e-16
NB	Hypernyms	2.646E-04	0.0000	< 2.2e-16
NB	Wiki words	3.625E-05	0.0000	< 2.2e-16
NB	clean	***6.261E-02***	0.0000	< 2.2e-16
NB	Bigrams	< 2.2e-16	0.0000	< 2.2e-16
NB	Wiki phrases	1.904E-02	0.0000	< 2.2e-16
NB	Wiki Bigrams	4.718E-03	0.0000	3.846E-09
NB	original	3.165E-02	0.0000	< 2.2e-16
NB	Hyponyms	< 2.2e-16	0.0000	< 2.2e-16
SVM	Lemmatised	6.075E-03	0.0000	< 2.2e-16
SVM	Synonyms	< 2.2e-16	0.0000	< 2.2e-16
SVM	Hypernyms	< 2.2e-16	0.0000	< 2.2e-16
SVM	Wiki words	< 2.2e-16	0.0000	< 2.2e-16
SVM	clean	6.918E-06	0.0000	< 2.2e-16
SVM	Bigrams	< 2.2e-16	0.0000	< 2.2e-16
SVM	Wiki phrases	< 2.2e-16	0.0000	< 2.2e-16
SVM	Wiki Bigrams	< 2.2e-16	0.0000	< 2.2e-16
SVM	original	< 2.2e-16	0.0000	< 2.2e-16
SVM	Hyponyms	< 2.2e-16	0.0000	< 2.2e-16
LSA	Lemmatised	2.537E-12	0.0000	< 2.2e-16
LSA	Synonyms	< 2.2e-16	0.0000	< 2.2e-16
LSA	Hypernyms	< 2.2e-16	0.0000	< 2.2e-16
LSA	Wiki words	< 2.2e-16	0.0000	< 2.2e-16
LSA	clean	2.977E-08	0.0000	< 2.2e-16
LSA	Bigrams	< 2.2e-16	0.0000	< 2.2e-16
LSA	Wiki phrases	< 2.2e-16	0.0000	< 2.2e-16
LSA	Wiki Bigrams	< 2.2e-16	0.0000	< 2.2e-16
LSA	original	< 2.2e-16	0.0000	< 2.2e-16
LSA	Hyponyms	< 2.2e-16	0.0000	< 2.2e-16

5.4. Post-Hoc Group Testing

Analysis of Variance testing constitutes what is known as an 'omnibus' test. Omnibus tests can discern that some significant difference exists between at least two of the categories or groups under inspection, but it is unable to specify which data points differ from which.

In order to differentiate between individual data points or, to look at things from a slightly different perspective, in order to group indistinguishable points together, post hoc testing is required.

Post-hoc tests are not included in the WRS2 R package used to conduct the earlier ANOVA testing, but they are available through the R package 'Psych' developed by Revele (Revele, 2015). The post-hoc procedures in the Psych package were used to group both message lengths and enhancements for all classifiers at a confidence level of 95%. Results are presented below.

The post-hoc grouping step assigns items to groups based on their mutual indistinguishability. For example, if observation 1 is not significantly different than observation 2 at the selected confidence level, then 1 and 2 will be assigned to the same group. However, if a third observation, 3, is significantly different from observation 1, but not significantly different from observation 2 then a new group will be formed containing observations 2 and 3. In this example observation 2 has membership of both groups, but observations 1 and 3 are each members of only a single, non-mutual group respectively. The tests for statistical differences, used in post-hoc group testing, are adjusted to take account of 'family-wise error' accumulation, which precludes the use of a repeated t-test, in much the same way as ANOVA avoids the same problem.

Table 5-5 Post-hoc Groupings of Enhancements for Naïve-Bayes Classification

Enhancement	Group	a	b	c	d	e	f	g	h	i
Bigrams	a	**a**								
Clean	b		**b**							
Hypernyms	c			**c**						
Hyponyms	d				**d**					
Lemmatised	b		**b**							
Original	e					**e**				
Synonyms	f						**f**			
WikiBigrams	g							**g**		
WikiPhrases	f						**f**			
WikiWords	c			**c**						

From the post-hoc groupings given above, it can be seen that, statistically, there is no clear difference (at 95% confidence) between the F1 Score performance of the Clean and Lemmatized enhancements (Group b). This is also true of the wordnet Synonyms and the Wiki Phrases enhancements (Group f) and of the wordnet Hypernyms and the WikiWords enhancements (Group c). This mirrors what might be understood, informally and graphically, from Figure 4-3 in which all three of these pairings show a close correspondence.

Table 5-6 Post-hoc Groupings of Text Lengths for Naïve-Bayes Classification

Length	Group	a	b	c	d	e	f	g	h	i
10	a	**a**								
15	b		**b**							
20	c			**c**						
25	c			**c**						
30	d				**d**					
35	e					**e**				
40	ef					**e**	**f**			
45	fg						**f**	**g**		
50	bg		**b**					**g**		
80	bg		**b**					**g**		
110	bg		**b**					**g**		
138	bg		**b**					**g**		

Once again echoing what may be intuitively observed on Figure 4-3, the post-hoc groupings for the text lengths indicate that no significant changes occur as text lengths decrease from 138 characters to 45 or 50 characters but, as the length further decreases, statistically significant changes in F1 Score performance become apparent.

Table 5-7 Post-hoc Groupings of Enhancements for SVM Classification

Enhancement	Group	a	b	c	d	e	f	g	h	i
Bigrams	a	**a**								
Clean	b		**b**							
Hypernyms	c			**c**						
Hyponyms	d				**d**					
Lemmatised	e					**e**				
Original	a	**a**								
Synonyms	f						**f**			
WikiBigrams	g							**g**		
WikiPhrases	h								**h**	
WikiWords	i									**i**

Once again, post-hoc groupings indicate that significant differences exist between most enhancement methods but the Bigrams and Original methods (Group a) remain indistinguishable.

Table 5-8 Post-hoc Groupings of Text Lengths for SVM Classification

Length	Group	a	b	c	d	e	f	g	h	i
10	a	**a**								
15	b		**b**							
20	c			**c**						
25	d				**d**					
30	e					**e**				
35	e					**e**				
40	e					**e**				
45	e					**e**				
50	f						**f**			
80	f						**f**			
110	b		**b**							
138	b		**b**							

Post-hoc grouping indications for SVM text lengths broadly follow the naïve-Bayes groupings, showing slight differences in performance at longer text length, but sudden decline in performance is delayed until approximately 30-character message length is reached.

Table 5-9 Post-hoc Groupings of Enhancements for LSA / SVM Classification

Enhancement	Group	a	b	c	d	e	f	g	h	i
Bigrams	a	**a**								
Clean	a	**a**								
Hypernyms	b		**b**							
Hyponyms	c			**c**						
Lemmatised	a	**a**								
Original	a	**a**								
Synonyms	d				**d**					
WikiBigrams	d				**d**					
WikiPhrases	b		**b**							
WikiWords	b		**b**							

Once again, post-hoc testing agrees with what may be visually ascertained from Figure 4-5. Although significant differences do appear between sets of enhancements, the differentiation is not as marked as with naïve-Bayes or SVM. This result must be taken in the context of a weaker F1 Score performance overall, and it may be thought of as a case of 'all enhancements performing poorly'.

Table 5-10 Post-hoc Groupings of Text Lengths for LSA / SVM Classification

Length	Group	a	b	c	d	e	f	g	h	i	j	k
10	a	**a**										
15	b		**b**									
20	c			**c**								
25	d				**d**							
30	e					**e**						
35	f						**f**					
40	g							**g**				
45	h								**h**			
50	b		**b**									
80	i									**i**		
110	j										**j**	
138	k											**k**

Post-hoc text-length grouping for LSA shows a marked differentiation between lengths. What is not readily apparent from the post hoc analysis, but which is very obvious on the graphical output in Figure 4-5, is that there is a pronounced decline in performance for all enhancements as text length increases beyond 20 characters.

5.5. Summary of Analysis and Evaluation

ANOVA and post-hoc testing of the results broadly confirms what might be expected after a visual inspection of the graphical presentation of the results data from Chapter 4.

Statistically significant results from both 2-way and 1-way robust ANOVA testing presented in Table 5-1, Table 5-2, Table 5-3 and Table 5-4 indicate that the performance of classifiers do indeed change as text length decreases (at 95% confidence) and provide sufficient evidence to reject, for all enhancements and classifiers, Null Hypothesis 1: *The performance of short-text classification enhancement methods, as measured by weighted average accuracy (F1 Score), will not change as target text length decreases.*

Statistically significant results in the interaction between enhancement method and text length in the 2-way robust ANOVA testing presented in Table 5-1, Table 5-2 and Table 5-3 indicate that the length related change in classification performance is related to the enhancement method in use and provide sufficient evidence to reject, for all enhancements and classifiers, Null Hypothesis 2: *The changes in performance of different short-text classification enhancement methods as target text length decreases, measured by weighted average accuracy (F1 Score), will not differ between enhancement methods.*

The research question *"Do the changes in performance of different short-text classification methods, as measured by weighted average accuracy (F1 Score), differ between text enhancement methods and classifiers as target text length decreases?"* may now be answered: The performance changes in all classifiers differ between all enhancement methods as target text length decreases.

The marked differences in the performance of additive and non-additive enhancements discussed in section 4.4, and the decline in performance of additive enhancements with increasing text length, may indicate that the additive component of these enhancements was not sufficiently specific: the non-specific additions may have effectively added noise rather than adding information. This conjecture may be supported by the observation that the Wiki

Phrases enhancement consistently out-performed the Wiki Words enhancement. These enhancements contain exactly the same texts, but in the case of Wiki Phrases the information is in the form of n-grams, treated by the classifiers as single units, whereas Wiki Words presents the information in the form of single words – an inherently less specific presentation of the same information. This observation may support the intuition that the addition of 'too many words' may effectively mask, rather than enhance, any signal: an intuition that motivated the work of Sun (2012). Further, since longer texts will generate more additions, any such effect might be expected to be more prevalent at longer text lengths, which matches the experimental results in which performance was seen to decline with increasing length.

Examination of appendices 8.2 and 8.3 might lead an observer to question the exact relevance or specificity of some of the enhanced texts: for example, the phrase "sound good" generates additive enhancements "*Epistemology_of_science"* and *"Headlands of South Africa"* amongst others. While it is true that potentially useful supplementations do not necessarily have to conform to common sense, and therefore may not appear obvious to a human observer, as implied by Pang, Lee and Vaithyanathan (2002), they must, in order to serve a useful purpose from a classification point of view, be repeatable and discriminatory. The risk is that such vague conceptual links may be so tenuous as to be unrepeatable for similar original messages, or so general as to be universally repeatable and therefore of no discriminatory purpose.

In order to tentatively test the hypothesis that over-supplementation may be detrimental to classification accuracy, a count was made of the number of additional, indivisible, words that each enhancement had added to each of the sixty thousand tweets (i.e. each n-gram was counted as a single word). Each of these counts was then divided by the word count of the original tweet to give the size of the addition expressed as a multiple of the original tweet size, which can be thought of as the 'additive footprint' of that enhancement on that text. The mean of this additive footprint was calculated for each enhancement and word length combination. The resulting, mean, 'additive footprints' are tabulated below.

Table 5-11 Mean Additive Footprints of Enhancements by Text Lengths

	Enhancement									
Text Length	original	clean	Lemmatised	Wiki words	Wiki phrases	Synonyms	Hyponyms	Hypernyms	Bigrams	Wiki Bigrams
10	1.0	0.8	0.8	53.7	19.9	8.0	24.3	8.7	0.1	1.3
15	1.0	0.7	0.7	58.5	21.4	7.6	21.0	7.7	0.2	1.2
20	1.0	0.7	0.7	59.7	21.4	7.2	18.8	7.0	0.2	1.5
25	1.0	0.7	0.7	56.7	20.1	7.2	17.7	6.5	0.2	1.3
30	1.0	0.7	0.7	55.4	19.8	6.9	18.2	6.7	0.2	1.2
35	1.0	0.7	0.7	54.7	19.5	6.9	18.0	6.7	0.3	1.2
40	1.0	0.6	0.6	54.7	19.5	7.0	18.4	6.8	0.3	1.3
45	1.0	0.6	0.6	53.9	19.2	7.0	18.3	6.8	0.3	1.2
50	1.0	0.6	0.6	55.1	19.6	7.0	18.8	6.9	0.3	1.2
80	1.0	0.6	0.6	55.7	19.8	7.0	19.3	7.0	0.3	1.1
110	1.0	0.6	0.6	53.8	19.2	6.8	18.6	6.7	0.4	1.1
138	1.0	0.6	0.6	53.1	19.0	6.7	18.3	6.6	0.4	0.9
Mean	**1**	**0.7**	**0.7**	**55.5**	**19.9**	**7.1**	**19.2**	**7.0**	**0.3**	**1.2**

From Table 5-11 it can be seen that the additive footprint for a given enhancement is relatively stable across the range of text lengths: however, it should be noted that the footprint is a proportion, so, for example, if the Synonyms enhancement adds 70 words to a ten-word text on average, a twenty-word text would, on average, have 140 words added.

The relatively constant size of the footprint across text lengths was confirmed by 2-way ANOVA testing, which indicates that within an enhancement there is no significant difference in footprint due to changing length. ANOVA results are presented in Figure 5-4.

Figure 5-4 ANOVA of Additive Footprints

```
Anova Table (Type III tests)

Response: footprint
                    Sum Sq     Df  F value    Pr(>F)
(Intercept)           4997      1   19.254 1.145e-05 ***
enhancement         13014025      9 5571.557 < 2.2e-16 ***
Length                 0     11    0.000         1
enhancement:Length    371063     99   14.442 < 2.2e-16 ***
Residuals          157456049 606690
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This allows the mean footprint to be used representatively of the entire enhancement over the range of text lengths. The mean additive footprints can now be assigned a rank, from 1, having the least additive impact, to 10, having the most additive impact.

Table 5-12 Ranked Results for Additive Footprints

	Enhancement									
Text Length	original	clean	Lemmatised	Wiki words	Wiki phrases	Synonyms	Hyponyms	Hypernyms	Bigrams	Wiki Bigrams
Footprint	**1**	**0.7**	**0.7**	**55.5**	**19.9**	**7.1**	**19.2**	**7.0**	**0.3**	**1.2**
Rank	4	2.5	2.5	10	9	7	8	6	1	5

Table 4-9, reproduced below, can also be transformed into a ranked result shown as Table 5-14

Table 5-13 Reproduction of Table 4-9 Areas under the Performance Curve for Enhancements and Classifiers

	Enhancement									
Classifier	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words
NB	11.000	10.632	9.421	9.176	10.610	10.859	9.695	10.271	9.761	9.411
SVM	10.696	10.645	9.690	9.266	10.611	10.836	9.967	9.920	9.274	8.636
LSA	9.047	9.017	8.828	8.479	9.024	9.038	9.048	8.914	8.845	8.764

Table 5-14 Ranked Scores for F1 Score Areas under the Performance Curve

Classifier	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words
NB	1	3	8	10	4	2	7	5	6	9
SVM	2	3	7	9	4	1	5	6	8	10
LSA	2	5	8	10	4	3	1	6	7	9

Combining Table 5-12 and Table 5-14, and then correlation-testing the F1 Score accuracy ranks against the additive footprint ranks using Spearman's rank-order co-efficient and calculating the associated z-score (Zar, 1972) gives Table 5-15

Table 5-15 Correlation of F1 Score Ranks and Additive Footprint Ranks

Ranks	Enhancement											
Classifier	Bigrams	clean	Hypernyms	Hyponyms	Lemmatised	original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words	Spearman's r	z-score
NB	1	3	8	10	4	2	7	5	6	9	0.851	2.55
SVM	2	3	7	9	4	1	5	6	8	10	0.875	2.63
LSA	2	5	8	10	4	3	1	6	7	9	0.632	1.90
Additive Footprint	1	2.5	6	8	2.5	4	7	5	9	10		

The values of Spearman's test indicate a strong correlation between increasing additive footprint and decreasing accuracy as measured by F1 score for the naïve-Bayes and SVM classifiers, and a moderate correlation for the LSA classifier. In all three cases, the one-tailed z-score indicates a significant correlation between increasing additive footprint and decreasing accuracy at the 95% confidence level.

This empirical result would suggest that enhancements which over-supplement the original text are likely to be counter-productive in terms of accurate classification, and that the greater the degree of over-supplementation the greater the negative impact on classification accuracy.

6. CONCLUSION

Chapter 6 will conclude the main body of this work, through the evaluation of the design, results and conclusions presented in earlier chapters, and will discuss gaps and opportunities for refinements and further work.

6.1. Research Overview

This research characterised the performance of the classification of short texts, using selected text enhancement methods and three separate classifiers, as a function of the short text length.

The average performance of three well-known classification methods, naïve-Bayes, Support Vector Machine and Latent Semantic Analysis (followed by Support Vector Machine) was measured for a binary classification task on short texts (tweets) of twelve different lengths, having 10 different enhancements applied to the texts before classification. The results of these tests were statistically analysed to ascertain whether performance differed between applied enhancement methods, whether performance differed as text-length decreased and whether there was any interaction between the length and enhancements with respect to classification performance. Some qualitative comparisons of text enhancement methods were also undertaken.

6.2. Problem Definition

The research problem was defined by the question:

"Do the changes in performance of different short-text classification methods, as measured by weighted average accuracy (F1 Score), differ between text enhancement methods and classifiers as target text length decreases?".

And by the two sub-questions:

"Does classification performance for any of the included text enhancement methods change as texts decrease in length?"

and

"If classification performance varies, as texts decrease in length, for two or more of the included text enhancement methods, do those performance variations differ between enhancement methods?"

6.2.1 Hypotheses

The primary purpose of the research was to establish the validity of the following hypotheses:

Hypothesis 1:

The performance of short-text classification enhancement methods, as measured by weighted average accuracy (F1 Score), will not change as target text length decreases.

Hypothesis 2:

The changes in performance of different short-text classification enhancement methods as target text length decreases, measured by weighted average accuracy (F1 Score), will not differ between enhancement methods.

Of secondary interest were qualitative comparisons of the relative performance impact of the tested text enhancement techniques.

6.3. Design

From a data-centric point of view, the experimental design was strong due, in large part, to the very large size of the Sentiment140 corpus which facilitated the creation of large subsets of tweets (5000 tweets per set) of very specific lengths, having an exact balance of sentiment (exactly 50% positive and 50% negative). These large, balanced sets provide a firm statistical foundation on which the rest of the work was based.

In hindsight, the inclusion of three separate classifiers added complexity to the project both in terms of coding and execution, and in terms of communicating the salient results. On balance, this was a necessary evil: one important meta-result produced is that the behaviour of two of the three classifiers was similar. The performance character of the LSA classifier was surprisingly poor, but that may be accounted for by the observation that the LSA method is significantly more complex than the others and has more options available to allow fine-tuning. Such fine tuning was beyond the scope of this work: but nevertheless, despite a noticeable difference in performance, some features of the LSA classifications held broadly similar trends to those observed for SVM and naïve-Bayes methods. This knowledge should allow future

work to proceed using only one of the three classifiers; naïve-Bayes is likely the best candidate for a single classifier design as it easily implemented and analysed, and shows relatively strong performance in terms of both accuracy and variance.

Indirectly, the inclusion of three classifiers also impacted the completeness of the work. The extra work required by the inclusion of the three classifiers precluded any attempt to conduct an in-depth analysis of classifier performance for any of the three – for this reason the issue of the poor performance, against expectations, of the LSA classifier was not touched upon.

In the domain of Twitter, it is conventional to think of message length as measured by character count. This makes perfect sense from a telecommunications perspective but, from the perspective of this project the experimental work may have been improved if a decision had been taken to measure text length in terms of word counts. The advantages to a word count approach are three-fold. Firstly, all the classifiers in this project, and most classifiers in the wider field of natural language processing, operate at 'word-granularity' or higher; that is, the smallest basic unit for manipulation or analysis is a word rather than a character. Therefore, character length is actually functioning as an imprecise proxy for word count when comparing classifiers, and that imprecision introduces an unnecessary degree of uncertainty into the results. Secondly, from the perspective of interpretability of results, people tend to naturally think about message lengths in terms of words – for human generated messages, it is more natural, and informative, to speak, for example, of a four-word message rather than a twenty-five character message. Thirdly, the classification of natural language is very often linked with the activity of concept extraction, sometimes implicitly, as in the case of LSA, and sometimes explicitly, for example in the work of Gabrilovich and Markovitch (2006) and Wang, Wang, Li, and Wen (2014). Although the exact correspondence between words and concepts is a complex and long debated issue in the fields of cognitive science and philosophy, discussed for example in Malt, Ameel, Gennari, Imai, Saji, and Majid (2011), it is safe to say that, at least sometimes, words relate to concepts whereas standalone characters never do, and so, word count may be a more relevant measure than character count for natural language classification.

6.4. Evaluation & Results

Formal ANOVA testing, at a 95% confidence level, of each of the enhancement methods for each of the classifiers, followed by Jonckheere-Terpstra trend testing, indicated that text length played a statistically significant role in classification performance for all enhancement-

classifier combinations. This result provides enough evidence to reject Hypothesis 1 and accept the alternative hypothesis:

The performance of short-text classification methods, as measured by weighted average accuracy (F1 Score), change as target text length decreases.

2-way ANOVA testing of text length and enhancement method indicated that there was a significant statistical interaction between text length and enhancement method influencing classifier performance. This result provides enough evidence to reject Hypothesis 2 and accept the alternative hypothesis:

Changes in performance of short-text classification methods, as target text length decreases, differ between text enhancement methods.

Less formal findings indicate that non-additive enhancement methods out-performed the more sophisticated 'concept mining' methods based on Wordnet and Wikipedia, and that the naïve-Bayes classifier out-performed the others, narrowly beating SVM, both in terms of the F1 accuracy score and in terms of the variance of its results.

The performance of the additive enhancements was somewhat disappointing, and may, as suggested in section 5.5, have been due to a lack of specificity and relevance in the added words and phrases.

6.5. Contributions

The primary contribution of this work is to have provided direct quantitative experimental evidence that classification accuracy declines with text length for non-additive text enhancements, and that the exact quantitative nature of that decline is dependent upon the enhancement or pre-treatment applied to the text and to the classifier in use.

While quantitative statistical differences do exist between enhancements, the qualitative changes in accuracy can be seen to start as text length decreases towards 50 characters for all non-additive enhancements, and become very pronounced below 20 characters for all variants of a message. This suggests that in the cases of naïve-Bayes and SVM classifiers, text might be usefully, if subjectively, considered 'short' at lengths below 80 characters and 'very short' at lengths of less than 20 characters.

It has been demonstrated that the naïve-Bayes classifier, along with simple non-additive text enhancements, such as bigram inclusion, provides a strong baseline against which to measure the performance of other, more sophisticated, classification mechanisms.

Some evidence, which may form the starting point for future work, has been provided to suggest that additive enhancement methods, without careful control, may overwhelm any actual signal present in the text though the addition of noise associated with poorly matched textual supplementation. This is circumstantially borne out by the warnings in Gabrilovich and Markovitch (2006) and their direct reference to an unspecified 'ablation' process to reduce noise, and by the 'pruning' quote form Sriram, Fuhry, Demir, Ferhatosmanoglu and Demirbas (2010) presented in section 2.2, earlier in this document.

This work has introduced the concept of 'additive footprint' to label the proportional increase in word count imposed upon a text by a given enhancement, and demonstrated that the additive footprint remains relatively constant for a given enhancement over a range of text lengths.

6.6. Future Work & Recommendations

This work gives rise to several avenues of possible continuation: some specific to this project and some of a more general nature.

In general, it would be useful to explore the exact nature of the text-length response of the additive enhancement methods. Using the data sets from this work, a baseline count of the mean number of words added by each enhancement for each text-length could be established. This number could be used to create 'control sets' of texts that had been enhanced with the appropriate number of a) random strings b) random dictionary words. These sets could then be used as a baseline to measure performance of the additive enhancements and to determine by exactly what margin the additive methods are better than random message extension.

Another possible avenue for additive enhancement methods is experimentation with part-of-speech filtering, either at generation time (e.g. send only adjectives to wordnet for supplementation) or at application time (e.g. accept only adjectives as supplemental words) or both together (e.g. supplement adjectives only with adjectives, supplement nouns only with nouns and so on.) Such a filtering mechanism could be potentially used to attempt to limit the addition of non-relevant words to the original text, complementing the work of Mertiya and Singh (2016) or Kamps and Marx (2002).

Use of a term-frequency/inverse-document-frequency step during the learning phase of a model could be used to generate a dictionary of words, from training data, which may be high potential candidate words for supplementation with Wikipedia or Wordnet: supplementing only on high potential words may reduce unwanted noise.

Future work specific to this work might usefully investigate the 'bump' in accuracy seen for many enhancement-classifier combinations at message lengths of 20 to 25 characters. Some preliminary investigation was carried out to rule out any peculiarity or data artefact that may cause this small increase in accuracy, but replacement of the original data sets had no effect. Remaining possibilities are that some internal feature of the classification algorithms, common to all algorithms, may be responsible or that there may be human-caused effect. A carefully designed experiment, which measures the classification accuracy for originally longer messages which have had some words excised before classification, would allow a comparison between, for example, texts that were originally 20 characters long, (and retain all of the original author-created context and structure), and texts which are 20 characters long after having had some fraction of their words artificially and randomly removed, (which would have a substantially impaired version of the original author-created context and structure). This comparison would highlight the effect of author-created context and structure, and if carried out over different text lengths, such an experiment may be able to determine whether author-created context and structure varies with text length: for example, it may (or may not) indicate that texts in the 20 to 25-character range have a higher degree of author-created context and structure, which might, tentatively, be attributed to an author's avoidance of ambiguity when composing shorter messages.

The prospect of further work on bigrams, and their extension into n-grams, has initial appeal as the bigram enhancement was the strongest of all enhancements tested. However, n-grams are, by their very nature, of limited utility as the text length decreases and the size of the n-gram increases because, for a message of word length X, there are only (X-n) +1 n-grams available.

The implementation of an ensemble of enhancements, rather than classifiers, either boosted or bagged, may prove to be an interesting implementation with which to experiment.

The final, and perhaps most interesting, unanswered question to be suggested for future work relates to the surprisingly poor performance of the LSA classifier. This classifier was the only one that received any tuning during the current project and, in theory, it should deal well with

any noise introduced through over-supplementation by additive enhancements. An in-depth investigation into the mechanisms underlying this poor performance would be enlightening.

Pragmatically, the possibility of real-world application of any of the additive enhancements, as configured in this project, must be viewed sceptically. Considerable time and effort is required to assemble the enhanced data sets, and processing time for classification is significantly longer than for the non-additive enhancements. Given that there was actually an overall decrease in performance for the additive enhancements, they cannot be recommended.

It must be restated however, that no effort was made in this work to tune either classifiers or enhancements, and the possibility most certainly exists that refinements or tuning could result in improved performance. If refinements of the additive enhancements could be found that would guarantee an increase in classifier performance, even at only some text lengths, then a situation-specific decision could be taken as to whether, in any particular circumstances, the trade-off between additional preparation and classification time was advantageous in light of any potential performance increases.

7. BIBLIOGRAPHY

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009, May). A study on similarity and relatedness using distributional and wordnet-based approaches. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 19-27). Association for Computational Linguistics.

Banerjee, S., Ramanathan, K., & Gupta, A. (2007, July). Clustering short texts using wikipedia. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 787-788). ACM.

Bellegarda, J. R. (2000). Exploiting latent semantic information in statistical language modeling. Proceedings of the IEEE, 88(8), 1279-1296.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. www, 7, 757-766.

Cai, L., & Hofmann, T. (2003, July). Text categorization by boosting automatically extracted concepts. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 182-189). ACM.

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391.

Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. Educational and Psychological Measurement, 34(4), 789-799.

Gabrilovich, E., & Markovitch, S. (2006, July). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In AAAI (Vol. 6, pp. 1301-1306).

Gabrilovich, E., & Markovitch, S. (2007, January). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In IJcAI (Vol. 7, pp. 1606-1611).

Genc, Y., Mason, W. A., & Nickerson, J. V. (2013). Classifying Short Messages using Collaborative Knowledge Bases: Reading Wikipedia to Understand Twitter. In # MSM (pp. 50-53).

Genc, Y., Sakamoto, Y., & Nickerson, J. V. (2011, July). Discovering context: classifying tweets through a semantic transform based on wikipedia. In International Conference on Foundations of Augmented Cognition (pp. 484-492). Springer Berlin Heidelberg.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. Machine learning, 11(1), 63-90.

Hu, X., Sun, N., Zhang, C., & Chua, T. S. (2009, November). Exploiting internal and external semantics for the clustering of short texts using world knowledge. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 919-928). ACM.

Imran, M., Mitra,P., & Castillo,C. (2016). Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages.In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 1638 – 1643.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98, 137-142.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. Biometrika, 41(1/2), 133-145.

Kamps, J., Marx, M., Mokken, R. J., & de Rijke, M. (2001). Words with attitude (pp. 332-341). Institute for Logic, Language and Computation (ILLC), University of Amsterdam.

Keller, F., Lapata, M., & Ourioupina, O. (2002, July). Using the web to overcome data sparseness. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 230-237). Association for Computational Linguistics.

Kim, S. B., Han, K. S., Rim, H. C., & Myaeng, S. H. (2006). Some effective techniques for naïve Bayes text classification. IEEE transactions on knowledge and data engineering, 18(11), 1457-1466.

King, R. D., Feng, C., & Sutherland, A. (1995). Statlog: comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence an International Journal, 9(3), 289-333.

Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. Information Processing & Management, 42(1), 56-73.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In Proceedings of the 19th annual meeting of the Cognitive Science Society (pp. 412-417).

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

Lewis, D. D. (1998, April). Naïve (Bayes) at forty: The independence assumption in information retrieval. In European conference on machine learning (pp. 4-15). Springer Berlin Heidelberg.

Lim, T. S., Loh, W. Y., & Shih, Y. S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine learning, 40(3), 203-228.

Ma, C., Xu, W., Li, P., & Yan, Y. (2015, June). Distributional representations of words for short text classification. In Proceedings of NAACL-HLT (pp. 33-38).

Malt, B. C., Ameel, E., Gennari, S., Imai, M., Saji, N., & Majid, A. (2011). Do words reveal concepts?. In *The 33rd Annual Conference of the Cognitive Science Society [CogSci 2011]* (pp. 519-524). Cognitive Science Society.

Mertiya, M., & Singh, A. (2016, August). Combining naïve Bayes and adjective analysis for sentiment detection on Twitter. In Inventive Computation Technologies (ICICT), International Conference on (Vol. 2, pp. 1-6). IEEE.

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

Mullen, T., & Collier, N. (2004, July). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In EMNLP (Vol. 4, pp. 412-418).

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

Peng, F., & Schuurmans, D. (2003, April). Combining naïve Bayes and n-gram language models for text classification. In European Conference on Information Retrieval (pp. 335-350). Springer Berlin Heidelberg.

Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008, April). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of the 17th international conference on World Wide Web (pp. 91-100). ACM.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003, August). Tackling the poor assumptions of naïve Bayes text classifiers. In ICML (Vol. 3, pp. 616-623).

Revele, W. (2015). Procedures for Personality and Psychological Research. Northwestern University.

Song, G., Ye, Y., Du, X., Huang, X., & Bie, S. (2014). Short Text Classification: A Survey. Journal of Multimedia, 9(5), 635-643.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010, July). Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 841-842). ACM.

Sun, A. (2012, August). Short text classification using very few words. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 1145-1146). ACM.

Wang, B. K., Huang, Y. F., Yang, W. X., & Li, X. (2012). Short text classification based on strong feature thesaurus. Journal of Zhejiang University-Science C, 13(9), 649-659.

Wang, F., Wang, Z., Li, Z., & Wen, J. R. (2014, November). Concept-based short text classification and ranking. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (pp. 1069-1078). ACM.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (2004, November). Latent semantic analysis. In Proceedings of the 16th international joint conference on Artificial intelligence (pp. 1-14).

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: measures of central tendency. Psychological methods, 8(3), 254.

Xu, Q. S., & Liang, Y. Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*(1), 1-11.

Yuan, Q., Cong, G., & Thalmann, N. M. (2012, April). Enhancing naïve Bayes with various smoothing methods for short text classification. In Proceedings of the 21st International Conference on World Wide Web (pp. 645-646). ACM.

Zar, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. Journal of the American Statistical Association, 67(339), 578-580.

Zelikovitz, S., & Hirsh, H. (2000). Improving short text classification using unlabeled background knowledge to assess document similarity. In Proceedings of the seventeenth international conference on machine learning (Vol. 2000, pp. 1183-1190).

8. APPENDICES

8.1. Glossary

analysis of variance: a collection of statistical methods used to analyse differences between group means.

ANOVA: see 'analysis of variance'.

bag-of-words: the treatment of a text segment as a set of words - ignores word order, semantics and grammar, but retains multiplicity.

bigram: set of two consecutive words from within a text (see also 'n-gram').

DBpedia: a structured database of Wikipedia data and metadata (see also 'Wikipedia').

dimensionality reduction: the mathematical operation of reducing the number of dimensions of a vector space by combining dimensions along which present data is highly correlated.

emoticon: a short combination of ascii characters chosen to represent a human face. Often used in short texts to convey emotion.

F1 score: the weighted average of precision and recall (see also 'precision' & 'recall').

hypernym: a word whose meaning covers the meanings of more specific words (see also ' hyponym').

hyponym: a word whose meaning is covered by the meaning of a less specific word (see also ' hypernym').

latent semantic analysis: a natural language processing technique based on term frequency coupled with a dimension reduction step.

latent semantic indexing: see 'latent semantic analysis'.

lemmatization: the conversion of words to their lemma or 'dictionary form'.

LSA: see 'latent semantic analysis'.

LSI: Latent Semantic Indexing - see 'latent semantic analysis'.

metadata: data that describes other data, eg classification hierarchies.

naïve-Bayes: a supervised learning method, often used for classification tasks, based on Bayes law and the naïve assumption of independence among predictors.

NB: see 'naïve-Bayes'.

n-gram: set of n consecutive words within a text where n is an integer number (see also 'bigram').

NLTK: Natural Language Toolkit - a python library providing functionality for the analysis of natural language.

omnibus test: test which can discern that some significant statistical difference exists between at least two of three or more groups under inspection, but it is unable to specify which groups differ from which.

precision: the ratio of correctly predicted positive outcomes to the total predicted positive outcomes.

python: a general purpose, high level, computer programming language often used in scientific applications.

R: an open source programming language and software environment for statistical computing.

recall: the ratio of correctly predicted positive outcomes to the actual number of postive outcomes.

supervised learning: machine learning methods which require the provision of a set of correct examples (training data) on which the model can be trained prior to testing and eventual production use.

support vector machine: a supervised learning method often used for classification tasks.

SVM: see 'support vector machine'.

synonym: words having the same meaning as another.

taxonomic: relating to taxonomies or naming schemes.

test data: data provided to a classifier without the associated correct classification information. Used to 'test' supervised learning models.

training data: data and the associated correct classification data provided to a classifier. Used to 'train' supervised learning models.

tweet: short message of no more than 140 characters used on the 'Twitter' network messaging application. (see also 'Twitter').

Twitter: social networking and messaging application for the broadcast of short messages known as tweets.

Wikipedia: online encyclopaedia which includes a taxonomy for the classification of included articles.

Wordnet: a large semantically oriented lexical database of English.

8.2. DBpedia – Returned XML for the word “sound”

Results of the query:

http://lookup.dbpedia.org/api/search.asmx/KeywordSearch?QueryString=%22sound%22

in XML format.

```
<ArrayOfResult xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns="http://lookup
.dbpedia.org/">
<Result>
<Label>Soundtrack</Label>
<URI>http://dbpedia.org/resource/Soundtrack</URI>
<Description>
A soundtrack can be recorded music accompanying and synchronized to the
images of a motion picture, book, television program or video game; a
commercially released soundtrack album of music as featured in the
soundtrack of a film or TV show; or the physical area of a film that
contains the synchronized recorded sound.
</Description>
<Classes/>
<Categories>
<Category>
<Label>Film and video terminology</Label>
<URI>
http://dbpedia.org/resource/Category:Film_and_video_terminology
</URI>
</Category>
<Category>
<Label>Soundtracks</Label>
<URI>http://dbpedia.org/resource/Category:Soundtracks</URI>
</Category>
</Categories>
<Templates/>
<Redirects/>
<Refcount>3081</Refcount>
</Result>
<Result>
<Label>Synthesizer</Label>
<URI>http://dbpedia.org/resource/Synthesizer</URI>
<Description>
A sound synthesizer (often abbreviated as "synthesizer" or "synth") is an
electronic instrument capable of producing a wide range of sounds.
Synthesizers may either imitate other instruments or generate new timbres.
They can be played (controlled) via a variety of different input devices
(including keyboards, music sequencers and instrument controllers).
Synthesizers generate electric signals (waveforms), and can finally be
converted to sound through the loudspeakers or headphones.
</Description>
<Classes/>
<Categories>
<Category>
<Label>Bass (sound)</Label>
<URI>http://dbpedia.org/resource/Category:Bass_(sound)</URI>
</Category>
<Category>
<Label>Keyboard instruments</Label>
<URI>
http://dbpedia.org/resource/Category:Keyboard_instruments
```

```
</URI>
</Category>
<Category>
<Label>Electronic musical instruments</Label>
<URI>
http://dbpedia.org/resource/Category:Electronic_musical_instruments
</URI>
</Category>
<Category>
<Label>New Wave music</Label>
<URI>
http://dbpedia.org/resource/Category:New_Wave_music
</URI>
</Category>
<Category>
<Label>Hip hop</Label>
<URI>http://dbpedia.org/resource/Category:Hip_hop</URI>
</Category>
<Category>
<Label>Contrabass instruments</Label>
<URI>
http://dbpedia.org/resource/Category:Contrabass_instruments
</URI>
</Category>
<Category>
<Label>Synthesizers</Label>
<URI>http://dbpedia.org/resource/Category:Synthesizers</URI>
</Category>
</Categories>
<Templates/>
<Redirects/>
<Refcount>2819</Refcount>
</Result>
<Result>
<Label>Motown</Label>
<URI>http://dbpedia.org/resource/Motown</URI>
<Description>
Motown is a record company originally founded by Berry Gordy, Jr. and
incorporated as Motown Record Corporation in Detroit, Michigan, United
States, on April 14, 1960. The name, a portmanteau of motor and town, is
also a nickname for Detroit. Now headquartered in New York City, Motown is
a subsidiary of The Island Def Jam Music Group, itself a subsidiary of the
French-owned Vivendi subsidiary, Universal Music Group.
</Description>
<Classes>
<Class>
<Label>organisation</Label>
<URI>http://dbpedia.org/ontology/Organisation</URI>
</Class>
<Class>
<Label>organization</Label>
<URI>http://schema.org/Organization</URI>
</Class>
<Class>
<Label>record label</Label>
<URI>http://dbpedia.org/ontology/RecordLabel</URI>
</Class>
<Class>
<Label>agent</Label>
<URI>http://dbpedia.org/ontology/Agent</URI>
</Class>
```

```
<Class>
<Label>owl#Thing</Label>
<URI>http://www.w3.org/2002/07/owl#Thing</URI>
</Class>
<Class>
<Label>company</Label>
<URI>http://dbpedia.org/ontology/Company</URI>
</Class>
</Classes>
<Categories>
<Category>
<Label>African-American history</Label>
<URI>
http://dbpedia.org/resource/Category:African-American_history
</URI>
</Category>
<Category>
<Label>Soul music record labels</Label>
<URI>
http://dbpedia.org/resource/Category:Soul_music_record_labels
</URI>
</Category>
<Category>
<Label>African-American culture</Label>
<URI>
http://dbpedia.org/resource/Category:African-American_culture
</URI>
</Category>
<Category>
<Label>Motown</Label>
<URI>http://dbpedia.org/resource/Category:Motown</URI>
</Category>
<Category>
<Label>Vivendi subsidiaries</Label>
<URI>
http://dbpedia.org/resource/Category:Vivendi_subsidiaries
</URI>
</Category>
<Category>
<Label>Record labels established in 1959</Label>
<URI>
http://dbpedia.org/resource/Category:Record_labels_established_in_1959
</URI>
</Category>
<Category>
<Label>Labels distributed by Universal Music Group</Label>
<URI>
http://dbpedia.org/resource/Category:Labels_distributed_by_Universal_Music_
Group
</URI>
</Category>
<Category>
<Label>Pop record labels</Label>
<URI>
http://dbpedia.org/resource/Category:Pop_record_labels
</URI>
</Category>
<Category>
<Label>Music of Detroit, Michigan</Label>
<URI>
http://dbpedia.org/resource/Category:Music_of_Detroit,_Michigan
```

```
</URI>
</Category>
<Category>
<Label>American record labels</Label>
<URI>
http://dbpedia.org/resource/Category:American_record_labels
</URI>
</Category>
<Category>
<Label>History of Detroit, Michigan</Label>
<URI>
http://dbpedia.org/resource/Category:History_of_Detroit,_Michigan
</URI>
</Category>
<Category>
<Label>Rhythm and blues record labels</Label>
<URI>
http://dbpedia.org/resource/Category:Rhythm_and_blues_record_labels
</URI>
</Category>
<Category>
<Label>Record labels established in 2011</Label>
<URI>
http://dbpedia.org/resource/Category:Record_labels_established_in_2011
</URI>
</Category>
<Category>
<Label>Companies based in New York City</Label>
<URI>
http://dbpedia.org/resource/Category:Companies_based_in_New_York_City
</URI>
</Category>
</Categories>
<Templates/>
<Redirects/>
<Refcount>2584</Refcount>
</Result>
<Result>
<Label>Sound recording and reproduction</Label>
<URI>
http://dbpedia.org/resource/Sound_recording_and_reproduction
</URI>
<Description>
Sound recording and reproduction is an electrical or mechanical inscription
and re-creation of sound waves, such as spoken voice, singing, instrumental
music, or sound effects. The two main classes of sound recording technology
are analog recording and digital recording.
</Description>
<Classes/>
<Categories>
<Category>
<Label>Audio engineering</Label>
<URI>
http://dbpedia.org/resource/Category:Audio_engineering
</URI>
</Category>
<Category>
<Label>Media technology</Label>
<URI>
http://dbpedia.org/resource/Category:Media_technology
</URI>
```

```
</Category>
<Category>
<Label>Sound production technology</Label>
<URI>
http://dbpedia.org/resource/Category:Sound_production_technology
</URI>
</Category>
<Category>
<Label>Sound recording</Label>
<URI>
http://dbpedia.org/resource/Category:Sound_recording
</URI>
</Category>
</Categories>
<Templates/>
<Redirects/>
<Refcount>2435</Refcount>
</Result>
<Result>
<Label>Sampling (music)</Label>
<URI>http://dbpedia.org/resource/Sampling_(music)</URI>
<Description>
In music, sampling is the act of taking a portion, or sample, of one sound
recording and reusing it as an instrument or a sound recording in a
different song or piece. Sampling was originally developed by experimental
musicians working with musique concrète and electroacoustic music, who
physically manipulated tape loops or vinyl records on a phonograph.
</Description>
<Classes/>
<Categories>
<Category>
<Label>DJing</Label>
<URI>http://dbpedia.org/resource/Category:DJing</URI>
</Category>
<Category>
<Label>Sampling</Label>
<URI>http://dbpedia.org/resource/Category:Sampling</URI>
</Category>
<Category>
<Label>Plagiarism controversies</Label>
<URI>
http://dbpedia.org/resource/Category:Plagiarism_controversies
</URI>
</Category>
</Categories>
<Templates/>
<Redirects/>
<Refcount>2028</Refcount>
</Result>
</ArrayOfResult>
```

8.3. Example enhancements of a single tweet

The example, below, illustrates the enhancements applied to a single tweet. The actual text passed to the classifier in each case is shown in italic typeface.

8.3.1 Original

The original text of the tweet as extracted from the Sentiment140 dataset.

@projectkpaz sounds good

8.3.2 Cleaned

The original text having punctuation and stop words removed, and twitter specific strings (e.g. hashtags, urls) replaced with standard tokens.

$mention$ sounds good

8.3.3 Lemmatised

The cleaned set (above) lemmatised using the NLTK python library.

$mention$ sound good

8.3.4 Bigrams

Enhanced by appending all bigrams from the lemmatised tweet back to the lemmatized tweet.

$mention$ sound good sound_good

8.3.5 Synonyms

Enhanced by appending all available wordnet synonyms for each word in the lemmatised tweet to the lemmatized tweet.

$mention$ sound good heavy beneficial estimable secure just unspoilt dear strait near respectable speech_sound full right goodness go ripe wakeless salutary auditory_sensation level-headed expert skillful in_force vocalize fathom legal dependable soundly honorable intelligent levelheaded good undecomposed proficient well-grounded safe vocalise phone unspoiled upright trade_good sound in_effect audio practiced effective commodity reasoned healthy adept profound well honest effectual skilful thoroughly serious voice

8.3.6 Hypernyms

Enhanced by appending all available wordnet hypernyms for each word in the lemmatised tweet back to the lemmatized tweet.

$mention$ sound good artefact occurrence pronounce sensation articulate denote body_of_water language_unit linguistic_unit measure seem quality enounce sense_impression appear natural_event sound_out sense_experience mechanical_phenomenon announce morality channel auditory_communication advantage vantage enunciate artifact water esthesis look occurrent cause_to_be_perceived happening sound_property quantify aesthesis sense_datum say

8.3.7 Hyponyms

Enhanced by appending all available wordnet hyponyms for each word in the lemmatised tweet back to the lemmatized tweet.

$mention$ sound good susurrus boom_out footstep blare zing clop honk clumping glug splosh blow trampling fungible kindness drumbeat glide tweet tinkle better benignity thump snap whack bombilation clank crash patter song beat burble rataplan muttering bang clang clopping whistle chirrup snarl consumer_goods strum reverberate tick dissonate desirableness merchandise worldly_good knock din bombilate rustle whish sonant shopping plunk ripple cry pealing chorus rub-a-dub commonweal importation throbbing whirr bong knell unison gargle fancy_goods worldly_possession vowel sigh twitter welfare product euphony whiz chime racket whir click popping clangor basic desirability birr benefit resonate pure_tone entrant ticking toll import speak chirk squelch noise worthiness wiseness twang bleep grumble summum_bonum step murmur sing exportation chink quack graciousness drum bombinate rumble salvage orinasal voiced_sound bombination vowel_sound narrow clump ticktock ding splash claxon tone thumping bell beneficence orinasal_phone knocking pop echo ultrasound gong thrum drygoods skirl ring paradiddle staple gurgle trump middling ting dissonance vibrato sporting_goods prepare tootle dub quaver clunking vibrate beep zizz music swoosh tintinnabulation bubble chug pitter-patter hum crack mutter wisdom toot tapping virtue jingle consonant ware chatter whirring pierce ticktack phoneme tink virtuousness moral_excellence future semivowel export voice benignancy vroom murmuring cackel tap ringing rattle resound dripping guggle thunk rap slosh common_good drum_roll pink purr thud pat waver optimum ping clunk babble trample murmuration boom racketiness soundness clink roll noisiness footfall play susurration slush buzz whistling whizz chirp peal make_noise clippety-clop rolling lap splat jangle swish clangour saintliness clip-clop twirp drip soft_goods drone click-clack swosh mussitation

8.3.8 Wiki Words

Enhanced by appending all available words in all the ‘labels’ contained in the top five Wikipedia hits for each word in the lemmatised text back to the lemmatised text.

$mention$ sound good Soundtrack Film and video terminology Soundtracks Synthesizer Bass (sound) Keyboard instruments Electronic musical instruments New Wave music Hip hop Contrabass instruments Synthesizers Motown organisation organization record label agent owl#Thing company African-American history Soul music record labels African-American culture Motown Vivendi subsidiaries Record labels established in 1959 Labels distributed by Universal Music Group Pop record labels Music of Detroit, Michigan American record labels History of Detroit, Michigan Rhythm and blues record labels Record labels established in 2011 Companies based in New York City Sound recording and reproduction Audio engineering Media technology Sound production technology Sound recording Sampling (music) DJing Sampling Plagiarism controversies Health Health Health promotion Personal life Good Morning America creative work owl#Thing work television show 1970s American television series Daytime Emmy Award for Outstanding Talk Show winners American news television series ABC News 1990s American television series Live television programs Radio programs on XM Radio 2010s American television series American Broadcasting Company network shows 1975 television series debuts English-language television series 2000s American television series 1980s American television series The Sydney Morning Herald written work newspaper creative work work periodical literature owl#Thing Newspapers published in Sydney Publications established in 1831 1831 establishments in Australia Cape of Good Hope Maritime history of South Africa Headlands of South Africa Headlands of the Western Cape Geography of Cape Town Face (professional wrestling) Professional wrestling slang

8.3.9 Wiki Phrases

Enhanced by appending all available ‘labels’, each treated as an indivisible string (n-gram), from the top five Wikipedia hits for each word in the lemmatised text, back to the lemmatised text.

$mention$ sound good Soundtrack Film_and_video_terminology Soundtracks Synthesizer Bass_(sound) Keyboard_instruments Electronic_musical_instruments New_Wave_music Hip_hop Contrabass_instruments Synthesizers Motown organisation organization

record_label agent owl#Thing company African-American_history Soul_music_record_labels African-American_culture Motown Vivendi_subsidiaries Record_labels_established_in_1959 Labels_distributed_by_Universal_Music_Group Pop_record_labels Music_of_Detroit,_Michigan American_record_labels History_of_Detroit,_Michigan Rhythm_and_blues_record_labels Record_labels_established_in_2011 Companies_based_in_New_York_City Sound_recording_and_reproduction Audio_engineering Media_technology Sound_production_technology Sound_recording Sampling_(music) DJing Sampling Plagiarism_controversies Health Health Health_promotion Personal_life Good_Morning_America creative_work owl#Thing work television_show 1970s_American_television_series Daytime_Emmy_Award_for_Outstanding_Talk_Show_winners American_news_television_series ABC_News 1990s_American_television_series Live_television_programs Radio_programs_on_XM_Radio 2010s_American_television_series American_Broadcasting_Company_network_shows 1975_television_series_debuts English-language_television_series 2000s_American_television_series 1980s_American_television_series The_Sydney_Morning_Herald written_work newspaper creative_work work periodical_literature owl#Thing Newspapers_published_in_Sydney Publications_established_in_1831 1831_establishments_in_Australia Cape_of_Good_Hope Maritime_history_of_South_Africa Headlands_of_South_Africa Headlands_of_the_Western_Cape Geography_of_Cape_Town Face_(professional_wrestling) Professional_wrestling_slang

8.3.10 Wiki Bigrams

Enhanced by appending all available ‘labels’, each treated as an indivisible string (n-gram), from the top five Wikipedia hits for each bigram in the lemmatised text back to the lemmatised text.

$mention$ sound good Verisimilitude Philosophical_problems Epistemology_of_science Realism Veracity