ISSN 1744-1986

Technical Report N° 2008/ 23

Evaluating semantic and rhetorical features to determine author attitude in tabloid newspaper articles

D. Foreman

28 June, 2008

Department of Computing
Faculty of Mathematics, Computing and Technology
The Open University

Walton Hall, Milton Keynes, MK7 6AA
United Kingdom

http://computing.open.ac.uk

Evaluating semantic and rhetorical features to determine author attitude in tabloid newspaper articles

A dissertation submitted in partial fulfilment
of the requirements for the Open University's
Master of Science Degree
in Computing for Commerce and Industry

Dugald Foreman
(W706595X)

10 March 2009

Word Count: **16315**

Preface

I would like to express my gratitude to the following people, all of whom have helped throughout this research project:

My tutor, Dr Paul Piwek for continuous useful advice through this project.

My mother, Lee Foreman, for giving me considerable amounts of her time as a second annotator for the newspaper texts and readily sharing difficulties about their interpretation.

Joanna Krupa, for her continued encouragement for me to stay the course.

Table of Contents

List of Equations

List of Figures

List of Tables

Abstract

This dissertation investigates the potential of families of machine learning features to improve the accuracy of a semantic orientation classifier that assesses attitudes of tabloid journalists towards the subjects of their opinion piece articles. A category of language, “language of judgement”, is defined by which a journalist expresses an opinion matching his overall opinion of an article's subject matter.

When the existence of "language of judgement" was investigated, high inter-annotator agreement on per-document author attitude was found (values of Fleisch and Cohen's kappa were both 0.845) along with moderate agreement on per-sentence classification of judgemental or non-judgemental language (Fleisch's kappa of 0.507 and Cohen's kappa of 0.499).

Three families of feature sets were defined to detect this language. The first family, “Semantic features”, motivated by consideration of theory of journalism, tags repetitions of nouns that are either located in particular sections of the article or occur multiple times in the article as potential language of judgement.

The second and third families, “rhetorical features”, draw on Mann and Thompson's Rhetorical Structure Theory. For the second family, rhetorical relations are tagged to indicate the presence of potential language of judgement. For the third family, rhetorical relations are considered to mark potential shifts into and out of language of judgement. Areas of articles between tags from the first family and tags from the second family are tagged with features from this third family, to indicate that the sentence is potentially within an area of language of judgement bounded by these rhetorical relations.

The feature sets were not very productive in acquiring judgemental language, together or separately. Precision of 0.405 for combined features was low but exceeded the overall percentage of judgemental language (32.8%). Recall of 0.162 was very low.

While experimentation with the testing corpus did not give strong evidence for value of the feature sets, cross-validation tests on the training corpus showed greater potential, achieving precision of 0.520 and recall of 0.200. Inspection of learning curves created with the training corpus for the combination of all features showed that learning of judgemental language was taking place. This was also true for the “rhetorical” second and third families when they were investigated separately but was not seen for the first family of features. Weaknesses in corpora construction methodology are considered potentially responsible for differences in results between corpora: suggested changes to remedy this, if more opinion piece articles can be collected, are described.

When classifying per-document author attitude, using human-annotated language of judgement was seen to improve the accuracy of a semantic orientation classifier that used Turney's PMI-IR algorithm (in comparison to use of all language in a document). However classification using language selected by the machine learning method did not lead to a similar improvement. The low precision and recall for acquisition of language of judgement obtained on testing corpus data is considered a likely cause of this.

Chapter 1 Introduction

1.1 Context of the research: semantic orientation

This dissertation investigates how a computer program can measure attitudes of tabloid journalists towards the subjects of their articles. Such measurement falls within the topic of determining semantic orientation. Determining semantic orientation can be seen as the extraction and classification of a text's emotional content. Research into semantic orientation takes place within a broader academic discourse of how computers can understand or induce human emotional states: affective computing. Computing applications are ultimately created in response to human needs; as Picard (1995, p. 14) states:

> "Emotions have a major impact on essential cognitive processes; neurological evidence indicates they are not a luxury [...] Computers that will interact naturally and intelligently with humans need the ability to at least recognize and express affect".

Turney (2007) lists examples of the use of technology based on research into semantic orientation. These includes both extraction of emotion (classification of product reviews as positive or negative) and creation of emotion (enhancement of chatbots to appropriately produce positive or negative responses). He also includes applications such as distinction of antonyms and synonyms: such semantic orientation technologies can serve as components within a larger system. The goal of determining author attitude pursued by this project is a form of emotion extraction.

The following sections present initial descriptions of areas of interest to this study in order to provide background to this dissertation's aims and research question. The

subsequent literature review further describes the relevance of these theories to the project's experimental design and their position within the context of academic discourse.

1.2 A definition of semantic orientation as Author Attitude

An initial definition of semantic orientation is found in Hatzivassiloglou and McKeown's (1997, p. 174) work on classification of the sentiment of individual adjectives, based on the work of Lehrer (1974):

> "The semantic orientation or polarity of a word indicates the direction the word deviates from the norm for its semantic group or lexical field"

This paper gives the example of the antonyms “hot” and “cold”, stating these words share the property of expressing temperature but have different orientations: these words thus exist on different ends of a hypothetical axis of meaning.

Following the work of Battistella (1990), these authors note that evaluative characteristics can be bound up in this semantic orientation. The evaluative dimension associated with the semantic orientation of a word depends on the particular semantic group that a word participates in.

Studies of semantic orientation vary in their chosen definition of this evaluative dimension (and also in the type of text under consideration). Early work by Spertus (1997) discusses a method of detecting flames - abusive messages posted in internet forums – and evaluates writer attitude along a dimension running from polite to abusive.

Turney's (2002) work on movie review classification uses an alternative evaluative dimension, placing reviews along an axis of "recommended" versus "not

recommended". Reviews recommending a movie lie on the positive end of a scale for this dimension whereas reviews critical of the movie lie towards the negative end of this scale.

Pang et al.'s (2002) work uses a less specific dimension of semantic orientation. Again movie reviews are considered. However these authors simply consider if an reviewer feels positively or negatively towards a movie: while this is compatible with Turney's axis, this dimension incorporates a broader range of feelings than recommendation alone. When comparing work on semantic orientation, one should remain aware that described results may refer to the pursuit of subtly different patterns of thought.

Indeed not all work on semantic orientation defines a specific axis: Wiebe et al. (2004) do not consider polarity (positive or negative orientation) but instead consider the higher-level question of classifying portions of a newspaper articles as subjective or factual. Based on these results, the authors then consider if these articles should be classified as opinion pieces.

With respect to the above studies, I consider author attitude, the specific dimension of semantic orientation used in this study, to incorporate a range of feelings as implied by the simple "positive" or "negative" of Pang et al. (2002).

However some restrictions must be placed on this dimension. Wiebe et al.'s (2004) aforementioned work notes that a piece of subjective language "expresses the subjectivity of a source, who may be the writer or someone mentioned in the text" (p. 280). Accordingly, care must be taken to ensure that an attitude found in a text genuinely belongs to a particular author as opposed to some participant in an article's story.

An example of such problematic language found in the corpus gathered for this project is:

> 'Gordon Brown led the salutes, hailing the White House ascent of Obama, 47, as “inspirational”' (Crick et al., 2008)

Although Brown's positive attitude towards Obama is described, the attitude of the writers towards Obama is not certain from this language in isolation. While it could be hypothesised that writers might selectively report the views of others to support their viewpoint, investigating such a hypothesis is out of scope for this dissertation. Views attributable to others will be deemed distinct from author attitude.

It also considered that writers are sincere in their viewpoints. While it may be to some extent valid that journalists write for specific audiences or simply to earn wages, stated author attitudes are taken at face value.

The nature of potentially relevant language to determine author attitude is now considered.

1.3 Relevant language for author attitude

In the process of designing an experiment for this study, I reviewed a large number of tabloid newspaper articles and created an informal taxonomy for classifying language within these articles. Table 1.1 describes these categories along with example language taken from an emotive right-wing opinion piece article "Bed-hoppers are screwing us all" (Shanahan, 2008a). This article presents the view that the British government is allowing Muslims to claim excess state benefits due to the possibility that a Muslim man may marry up to four times.

Category	Description	Example
Factual	Language making a statement that a human annotator would consider factual.	"Islam allows up to four wives."
Judgemental	Language making a statement that a human annotator would consider to express an emotion held by an author in a sense (positive or negative) matching the overall sense (positive or negative) of the article.	"TALK about screwing the poor old taxpayer"
Counter-argument	Language that a human annotator would consider to express an emotion (positive or negative) held by an author that is opposite to the overall sense (negative or positive) of the article. This does not preclude the possibility that such an emotion is being expressed for rhetorical purposes, to enhance the overall sentimental impact of the article.	"Muslim marital customs in their own lands are not my business."
Digression	Language seemingly unrelated to the main subject of an article. Such language is observed to frequently occur at the end of tabloid articles after discussion of the main subject is complete.	“WHY the fuss over an MP being bugged?"

Table 1.1: Model of language categories

This dissertation does not attempt to formally prove the taxonomy's existence. It is presented as background to the overall approach used to design an experiment assessing author attitude. Within the taxonomy's definitions, factual and judgement language are closest to the main subject of a text whereas language of counterargument or digression seem more likely to discuss other subjects.

Pang et al. (2002) state results for semantic orientation are generally worse than for established methods of topic classification: accurate determination of semantic orientation seems more difficult than topic determination. As explanation for this, these authors identify a phenomenon of "thwarted expectations ... where the author sets up a deliberate contrast to earlier discussion". Thwarting a reader's expectations seems likely to occur in emotional features contained in counterargument.

It is hypothesized that if language of judgement can be correctly identified, a computer program that examines such language alone will avoid this "thwarted expectations" phenomenon and be more accurate than a program that examines all the different types of language within a text. This seems intuitively correct given the definition stated above for language of judgement.

Investigating this hypothesis has two consequences for the research:

- Using all classes of language from the text will give a baseline against which improvements in classifier accuracy can be measured.
- An attempt at improvement in classifier accuracy will be made by attempting to find a particular subset of the language within a text, the language of judgement.

Neither of these consequences are novel when the existing discourse on determining semantic orientation is considered.

1.4 Relevant techniques for determining semantic orientation

This section undertakes this consideration to demonstrate this and introduce techniques used in the research.

1.4.1 Baseline method

Turney's (2001) method tags adjectives or adverbs in a movie review and then estimates the pointwise mutual information (PMI) shared between each tagged word and words with known positive and negative connotations. The underlying algorithm includes an information retrieval (PMI-IR) component which uses numbers of hits returned from internet search engine queries for words and combinations of words as input data. The value for the semantic orientation of the overall document is calculated as an average of the values of the semantic orientation of individual words.

The current research uses PMI-IR to calculate overall document sentiment in the same manner as Turney. This method is discussed further in the literature review section. Other methods for determining semantic orientation are now briefly discussed.

1.4.2 Methods for determining semantic orientation

Methods for determining semantic orientation described in the literature can be divided in two main classes:

- methods which analyse features found throughout the entirety of a document
- methods which select portions of the document considered to contain particularly useful features for determining sentiment and then analyse this reduced set of features

While the previously described method used by Turney (2001) is a simple example of the first approach, whole-document analysis may involve sophisticated processing. Pang et al. (2002) consider approaches looking at the whole document to calculate overall semantic orientation: Naïve Bayes, Maximum Entropy and Support Vector Machines

(SVMs). These authors concluded that Support Vector Machines perform slightly better than the alternatives.

The literature review discusses several methods involving selection of portions of a document. However the most relevant work for this research is Taboada and Voll (2007). This work compares two distinct approaches that attempt to focus on particular areas of a text and so avoid digressions by an author from his central topic. The first method is based on rhetorical structure theory (RST).

Mann and Thompson (1988) provided an early detailed peer-reviewed description of RST. Mann and Taboada (2005) indicate this article is now difficult to obtain and recommend the use of a more complete report, Mann and Thompson (1987), which was the basis for the peer-reviewed article.

This report introduces RST as "a descriptive theory... of the organisation of natural text... (which) ... identifies hierarchic structure in text" (p. 1). It describes the relations between text parts in functional terms. Relations such as CONTRAST, CONCESSION, ELABORATION or SUMMARY are introduced (CONCESSION and CONTRAST relations will be discussed in greater detail in the literature review).

RST analysis thus uncovers a hierarchic tree of relations where relations cover different spans of a text and a relation may be a subcomponent of a higher level relation. These relations connect two types of discourse spans. A nucleus is the span that is "more essential to the writer's purpose" (p. 31) and consequently may be more central to the meaning of the text under consideration.

Given this definition of a nucleus, Taboada and Voll's (2007) first approach extracts language held in nuclei from a text and applies a calculation similar to Turney's PMI-IR

method on the adjectives contained within this language to derive overall document sentiment.

Their second approach uses a machine learning method to extract on-topic sentences. In this approach a decision tree classifier is trained on sentences that have been annotated as on-topic by human annotators. The most common adjectives within a corpus are used as features to train the classifier. After training, the classifier attempts to find on-topic language within unannotated documents. The same calculation based on Turney's method is then applied to this language to derive overall document sentiment.

Discussion of the aims and objectives of the present work reveals similarities to both approaches used by Taboada and Voll.

1.5 Project aims and objectives

This project aims to create a classifier that can acquire language of judgement after training on a corpus of texts that have been annotated at the sentence level for language of judgement (using the previously described taxonomy). Since language of judgement was defined earlier as "matching the overall sense (positive or negative) of the article", annotators also have to decide on the overall direction of author attitude for these articles.

After training, the classifier will attempt to find judgemental language within a separate testing set of documents. This discovered language will be used to classify the documents using Turney's PMI-IR method. Potential overall gains in classifier accuracy will be measured by comparison against the previously-discussed Turney method baseline.

The use of a machine learning approach described above is similar to the work of Taboada and Voll. However both the definition and the intent of the features used to train the classifier differ. Taboada and Voll annotate sentences as on-topic (or off-topic) for particular review categories (such as books or cars) and use individual words as features to train separate models to identify on-topic language for each category. The present work seeks to identify language of judgement and defines a more abstract set of features by considering the thought processes of the journalists writing the articles under consideration.

As previously discussed, language of judgement seems closest to the main subject of a text in the mind of a journalist since it is neither digression or counterargument. While language close to the main subject may also be factual, the assumption made here is that such factual language may contain less sentiment-bearing features than other types of language: consequently if a filtering process acquires factual as well as judgemental language, final results for overall sentiment classification may still be acceptable.

To find language close to article subjects, it is hypothesised that particular areas of articles, chosen after review of material discussing theory of journalism, may contain nouns closest to the overall subject. It is further hypothesised that repetition of these nouns throughout the article is considered a potential indicator of closeness to the main subject and so of potential language of judgement. The literature review discusses issues around "the main subject" of a text. A subsequent research methods section gives details and justification for this first set of judgement-seeking features.

A further hypothesis is that the different types of rhetoric may act as signals for language of judgement. Precedent for considering the value of individual rhetorical relations in determining sentiment orientation is found in recent work by Taboada

(2007). This found useful results for assessment of semantic orientation when focusing on concessive relations (this work also briefly mentions future work assessing the presence of judgement with machine learning in the context of appraisal theory). In the current work, tagging sentences with the appropriate highest-level relation in the sentence's hierarchic tree creates a second set of judgement-seeking features which allow exploration of potential association between judgemental language and rhetorical relations.

Mann and Thompson (1987) state that text covered by a CONCESSION relation can be considered to embody a situation along with a potentially incompatible situation that is also considered true. Similarly they describe CONTRAST relations as covering two different situations that are similar in some aspects but differ in others. This project will use Marcu and Soricut's (2003) SPADE parser to discover relations: A limitation of SPADE is that it outputs internally represented CONCESSION, CONTRAST and ANTITHESIS relations as CONTRAST relations (seen by examination of source code for SPADE v0.9). While SPADE might be modified to output these internal representations, validity of results after modifying this software is unclear. The following discussion does consider CONCESSION and CONTRAST relations as broadly equivalent in how they hold argumentative language.

A key point here is that such relations cover multiple situations. It is thus considered that they may have higher potential to cause a shift in the current subject of the article than other types of relation. This shift may be accompanied by a shift in the type of language in use so causing a movement into or out of language of judgement.

The example in table 1.2 taken from the corpus show movement from language of judgement to another class of language (sentences that were annotated as language of judgement are preceded by an "=" sign).

Article Subject: John Cena
Overall orientation: Positive
=/Cena is a workhorse. =/He's a tireless promotional machine. And the project, event, DVD, pay per view, film, CD, and merchandise he promotes are all branded "WWE". =/There's not one single wrestler I've met in the past two decades with Cena's drive, ambition and determination to give every fibre of his existence to the company. /Triple H may have married into the 24/7 life of a McMahon Family member, but he likes to go home every now and then.

Table 1.2: Example 1 - 'Why Vince McMahon loves John Cena' (Heyman, 2008)

In the final sentence a move occurs away from language of judgement (language that is in the same sentiment direction as the overall article). This final sentence was found to participate in a CONTRAST relation when processed with the SPADE parser. As stated earlier CONTRAST relations cover two different situations that are similar in some aspects but differ in others. Hence this example shows the possibility that a shift might occur away from language of judgement when a CONTRAST relation is encountered.

An example of a shift into language of judgement on encountering language indicating a contrast is seen in the extract from a negatively-orientated article in table 1.3.

Article Subject: Britannia High
Overall orientation: Negative
Britannia High, which famously kicked off the TV Awards and its first episode singing: “This could be the start of something good.” =/But then proved to be the exact opposite. =/An all-singing, all-dancing, issue-driven Sylvia Dung horror show that comes with little or no pedigree, just lots and lots of Chum.

Table 1.3: Example 2 - 'Britannia High ... all mingin' all dancin'' (Ross, 2008b)

Determining if CONTRAST (or any other relations) are indeed of value in determining shifts into and out of language of judgement across a large corpus is initially an open question. As Biber et al.'s (2002) textbook on corpus linguistics states "with a large amount of language, it is ... difficult to keep track of multiple contextual factors" (p. 3). The value of different relations must be determined experimentally. Accordingly an additional aim of this project is to assess if relations can be used in such a way.

To assess this, a third set of judgement-seeking features is created to describe potential for a particular relation to begin or end an area of language of judgement. A precise definition of how such potential features can operate is described in section 3.3.3.

1.6 Overview

Chapter 2 of this dissertation reviews additional literature on semantic orientation and RST. It also presents some discussion of the theory of journalism. Chapter 3 reviews the key hypotheses of the research leading to statement of the research question. It then defines features associated with these hypotheses. Chapter 4 discusses the data

collection process by describing the methodology for creating and annotating the experimental corpus along with particular issues encountered during annotation specific to the texts under consideration. Chapter 5 discusses the project's methods for processing this data by describing the experimental steps followed to measure semantic orientation. Chapter 6 presents the results of this experiment, focusing on acquisition of language of judgement and overall document classification. Chapter 7 concludes the dissertation and presents learnings from the research.

Chapter 2 Literature Review

The literature review first considers the sentiment axis used to classify the overall document and then discusses the nature of lexical semantic features and how they can be quantified. It then considers appropriate features for classifying language of judgement from consideration of the practice of journalism and RST.

2.1 Sentiment axis for document classification

The following discussion on sentiment analysis further considers the sentiment axis on which documents will be classified by this research. It then considers work on semantic orientation at the word level and discusses:

- the nature of lexical features used for extracting sentiment.
- Turney's (2001) algorithm for extracting PMI.

2.1.1 Units of scale for measuring sentiment

The units of the scale on which document-level results are measured is another choice in studies of sentiment orientation. Turney (2001) uses a binary classification of positive or negative. In contrast, Pang and Lee (2005) use a multi-point scale allowing for quantification of discovered sentiment.

Since the goal of the research is to classify opinion pieces, a simple binary scale is used. This seems a natural choice - an opinion piece should express an author's opinion in one direction or other. This choice also allows easy comparison with other work such as Taboada and Voll (2007).

However this issue will be discussed further in section 4.1 as it was found to be problematic during corpus construction.

2.1.2 Influence of subject matter on direction of orientation

Determining subject matter of the tabloid articles is explicitly placed out of scope. While text summarisation techniques exist, the desire to focus on particular sets of semantic and rhetorical features precludes their use in a work of this size.

However, while subject matter is not explicitly considered, annotator interpretation of article subject matter can affect the overall direction of sentiment assigned. An example of this occurred during annotation of Burchill's (2008) article "Double standard hits Sienna's rep" which discusses an affair conducted by the celebrity Sienna Miller and states she is the victim of society's hypocritical views. In initial annotation trials, one annotator considered the subject to be Sienna herself and considered the author positive towards her as a subject. The other annotator considered the subject to be society's views and considered the author negative towards those views.

Another article, "Gordon will make JK's £1m vanish" (Shanahan, 2008b) could be considered to have three different potential subjects: the writer J.K. Rowling (who donated money to the labour party), the British Prime Minister Gordon Brown (who the author feels will waste the money) and the labour party itself. In addition to potential effects on overall article sentiment, the subject chosen will affect an annotators choice of how to categorise language. For example, the sentence

> "Author JK Rowling is an amazing lady, a brilliant writer and has done wonders for kids by getting them hooked on reading."

might be judgemental if J.K. Rowling is chosen as the subject. However if one of the other two potential subjects is chosen such language would likely be classified as digression.

This potential for multiple subjects seems more likely to be an issue for tabloid newspapers than for other targets of work on semantic orientation. For example Turney's (2002) study of movie reviews or Taboada and Voll's (2007) study of product reviews are likely to have a single movie or product as the subject. Multiple subjects may also be more of an issue for this dissertation because the explicit desire to consider author attitude requires a focus on a particular subject rather than the text as a whole.

2.1.3 Determining semantic orientation at word level

While Turney (2001) considered both adjectives and adverbs when calculating semantic orientation using PMI, this study will look at adjectives alone. The initial practical motivation for this results from use of Yahoo's WebSearch API (Yahoo, 2009). This API provides an interface to Yahoo's search engine that allows hits counts to be obtained for particular searches and was chosen as queries could be scripted without breaching the API's terms of use (other search APIs which the author could access were designed for web-based environments, precluding development of a tool-chain). The disadvantage of this API is that it is “limited to 5,000 queries per IP address per day” (Yahoo, 2009). This decision was motivated by the desire to minimise queries and justified by the use of adjectives alone in Taboada and Voll's (2007) work given that they consider that "adjectives...convey a high degree of opinion" (p. 337). This seems acceptable since Taboada and Voll is the more recent work and adverbial discourse markers will be used to detect cross-sentential relations in the course of the experiment.

However feature definition, for discovery of argument or sentiment, involves more than choice of part of speech. Wiebe et al.'s (2004) work on identification of sentiment-bearing features considers words that are hapax legomena (unique words) within a text as well as more complex features - collocations of higher precision than their component features and ugen-n-grams (sets of words found in proximity to unique words). Wiebe et al. also use distributional similarity to enlarge the feature set beyond annotated corpora through discovery of words that share mutual information with other words in the corpus. While these approaches might be of value, since this research aims to assess the value of semantic and rhetorical features in acquiring language of judgement, these methods of feature creation are placed out of scope.

Turney's (2001) method for calculating PMI is now summarised by reference to the key equations used to produce a value for the semantic orientation of a word along a particular axis.

$$\mathrm{PMI}(word_1, word_2) = \log_2 \left[\frac{\mathrm{p}(word_1 \,\&\, word_2)}{\mathrm{p}(word_1)\,\mathrm{p}(word_2)} \right] \quad (1)$$

Equation 2.1: PMI calculation from probabilities

Paraphrasing Turney, in equation 2.1 p(word1 & word2) represents the probability that the two words co-occur. If no statistical connection exists between the two words then the probability that they both occur is equal to p(word1)p(word2). Accordingly the quantity p(word1 & word2)/(p(word1)p(word2)) measures statistical dependence between the words and the log of this value represents the amount of information given for the presence of one of the words when the other is observed.

$$SO(phrase) = PMI(phrase, \text{“excellent”}) - PMI(phrase, \text{“poor”}) \quad (2)$$

Equation 2.2: Semantic orientation calculation from PMI

Turney uses pre-defined sets of words with positive and negative semantic orientation. Subtracting the PMI of some target word with respect to a negative word from the PMI of that word with respect to a positive word (equation 2.2) gives an overall value of semantic orientation in a positive or negative direction. Since the PMIs of both words are expressed as logarithmic functions, the subtraction operation can be rewritten as the log of the operand of the function to acquire the PMI of the positive word divided by the operand of the function to acquire the PMI of the negative word. When this is done, the terms for P(phrase) will cancel out.

The remaining probabilities can then be converted to numbers of hits returned from a search engine (the IR component of PMI-IR). This conversion can be done by observing that higher values of hits are proportional to higher probability that the word will occur in the document. The probability of a word’s occurrence is equal to the total number of hits divided by the number of documents indexed, the maximum possible hits. These hit counts can now be substituted into equation 2.2, giving an equation for calculation of word sentiment, equation 2.3.

$$\mathrm{SO}(\textit{phrase}) = \log_2 \left[\frac{\mathrm{hits}(\textit{phrase}\ \mathrm{NEAR}\ \text{“excellent”})\ \mathrm{hits}(\text{“poor”})}{\mathrm{hits}(\textit{phrase}\ \mathrm{NEAR}\ \text{“poor”})\ \mathrm{hits}(\text{“excellent”})} \right] \quad (3)$$

Equation 2.3: Semantic orientation calculation from hit counts

The term for the number of documents indexed cancels out in creation of equation 2.3. This is of practical importance as search engines companies (including Yahoo) do not provide up-to-date counts of documents indexed. Without this count, when this term does not cancel out (for example when only calculating PMI), this unknown term will be present.

The above equations do not represent the only approach available for measuring sentiment. Turney and Littman (2003) compare more sophisticated algorithms using PMI, the vector-based method of latent semantic analysis (LSA) and Hatzivassiloglou and McKeown's (1997) method of graph-clustering. Taboada and Voll (2007) found better results using a hand-ranked dictionary as opposed to PMI-IR and also found the changing nature of internet-generated data to be a source of instability for PMI-IR (Kilgarriff (2007) makes further criticisms of methods with an IR component). The above method is used in this project since it is a well-known technology within the field, is easy to implement and a common baseline is required as opposed to an optimal method.

2.2 Potential of journalistic practice for determining language of judgement

The starting point for feature design was the desire to consider the thought processes of the journalist whose attitude will be classified. This study considers that the practice of journalism has its own body of knowledge. The existence of publications such as the

Handbook of Independent Journalism (Potter, 2006) is evidence for this. Tabloid journalists are explicitly assumed to be professionals who follow practices from this body of knowledge to some extent.

Consequently considering the nature of the texts may provide useful insights for location of language related to the central subject matter (which as previously discussed, may be factual or judgemental). The aforementioned Handbook of Independent Journalism (Potter, 2006), a guide for writing newspaper articles, describes the concept of a lede (alternatively spelt lead) – an introductory section at the beginning of an article that introduces its subject. As an example, Shanahan's (2008c) article titled 'Bank bailouts and no one's bovvered [sic.]' has the lede:

> "WELL, congratulations. You are now the proud part-owner of TWO busted banks."

It is observed that nouns may be more useful for the purposes of this study than verbs or adjectives for identification of language close to the subject matter of a text. Justification for this observation comes from the field of topic segmentation, the discovery of story boundaries within a text, where nouns or noun phrases are frequently chosen as features of interest. Matveeva and Levow (2007) is a recent study in this area that divides potential features into two sets, "nouns and the rest of the vocabulary" (p. 352).

Accordingly, it is hypothesised that looking for sentences in an article where nouns found in a lede are repeated may allow access to areas of language related to the central subject matter: an example of such language found by looking for repetitions of "banks" in this article is:

> "The state is using billions of OUR money to buy up banks and there has been not a word of discussion."

This language was annotated as judgemental in nature: it is thought to relate to the central subject matter of the article. Section 3.2 describes how different types of features are defined for this research based on repetition of nouns from the lede and other sections of the article (these additional feature types will be derived through further reflection on journalistic practice).

Although Matveeva and Levow (2007) exclude proper nouns, these are considered relevant to this research. Many corpus articles are about celebrities or other personalities. Allowing access to potential subject matter language through repetition of these names seems appropriate.

An additional consideration is how to define the boundaries of an article lede. Potter(2006) gives the recommendation that “Each paragraph contains one main idea” (p. 24). Review of the training corpus showed a tendency to italicize the first paragraph of articles. This may be interpreted as a desire to emphasis this portion of the article (as might be done for a lede) or to mark it as a cohesive unit. Accordingly the first paragraph is defined as the lede. It is however noted that many articles in the corpus contain mostly single sentence paragraphs. In such cases, the lede is effectively one sentence long.

Further discussion of the nature of ledes will take place in the following chapter.

2.3 Rhetorical Structure Theory

The following section first presents theoretical background on RST, in particular describing relations that may describe multiple situations. As discussed in section 1.5, such relations may have greater potential to cause shifts into or out of judgemental language (section 3.3 will discuss how potential features to detect these shifts may be specified).

2.3.1 Discussion of RST Relations

The following section describes the CONCESSION and CONTRAST relations using definitions taken from Mann and Thompson's (1987) original work on RST. In addition to presenting key concepts of RST, it gives a theoretical basis for how these relations may hold different situations and so potentially indicate a shift in subject matter as discussed in section 1.5.

Figure 2.1 defines CONCESSION relations. A relation of this type connects two spans of discourse, a nucleus N and a satellite S. The abbreviation "W" in the above template indicates the writer, and "R" the reader. As the constraint on the N+S situation makes clear, two different situations are involved when such a relation occurs. Given that multiple situations are present, corresponding shifts in subject matter and type of language (moving into or out of judgemental language) may occur.

relation name:	CONCESSION
constraints on N:	W has positive regard for the situation presented in N;
constraints on S:	W is not claiming that the situation presented in S doesn't hold;
constraints on the N+S combination:	
W acknowledges a potential or apparent incompatibility between the situations presented in N and S; W regards the situations presented in N and S as compatible; recognising the compatibility between the situations presented in N and S increases R's positive regard for the situation presented in N	
the effect: R's positive regard for the situation presented in N is increased	

Figure 2.1: Definition of CONCESSION relation (Mann and Thompson, p. 15)

All relations within Mann and Thompson's initial definition of RST may be described with similar schema templates. Accordingly the template in figure 2.2 describes a CONTRAST relation.

relation name: CONTRAST
constraints on N: multi-nuclear
constraints on the combination of nuclei:
no more than two nuclei; the situations presented in these two nuclei are (a) comprehended as the same in many respects (b) comprehended as differing in a few respects and (c) compared with respect to one or more of these differences
the effect: R recognised the comparability and the difference(s) yielded by the comparison is being made
locus of the effect: multiple nuclei

Figure 2.2: Definition of CONTRAST relation (Mann and Thompson, p. 75)

A difference between the CONTRAST and CONCESSION relations is the presence of two nuclei in CONTRAST relations. These nuclei are however considered similar in function to the nucleus and satellite of a CONCESSION relation given both spans hold different situations. Again the constraint on the combination of nuclei implies two different situations are involved and shifts in subject matter and type of language may occur.

2.3.2 Miscellaneous issues in RST

Since RST's initial development, debate has continued around the definition of an appropriate set of relations. Knott's (1996) methodology for deriving relations attempted to tackle this: however his methodology has not been universally adopted. The set of

relations used in this project is determined pragmatically: relations output by the SPADE parser will be used.

A second ongoing debate concerns the hierarchical nature of the RST tree. RST is not the only theory that attempts to describe textual structure. Hobbs (1985) gives an example of incoherent but realistic discourse which is better represented by a graph of relations than a tree. Mann and Taboada (2007) acknowledge potential value in work by Wolf and Gibson (2005) on newspaper texts which they state argues that “more powerful data structures than trees are necessary to represent discourse structure”.

It can be argued that the tabloid articles may exhibit a degree of incoherence, at least at the highest structural level. Within the corpus collected for this study, the bulk of Blunkett's (2008) article "Crying wolf is a risky game" discusses the resignation of the politician David Davies in protest at Government security measures. However the final paragraphs of the article discuss issues related to the public health service then the article closes with a single unrelated sentence on the incomprehensibility of a European treaty.

A final issue in RST is that, as Mann and Thompson (1987) identify, RST analysis of a text may have more than one possible result: human annotators may disagree how to mark up an ambiguous text or a text may have multiple possible annotations. This can cause difficulties both in training and testing phases of research involving RST. For this work relations are obtained in an automated fashion through SPADE or by detecting discourse markers described in Taboada (2006), as described in section 3.3.1. Failure of these technologies to accurately classify relations due to this (or other errors) is accepted as a source of noise in experimental results.

Chapter 3 Research Question and Feature Definitions

This chapter reviews key hypotheses underlying the research then states the research question based on these hypotheses. It then describes procedures for defining families of machine learning features for language of judgement associated with these hypotheses.

3.1 Research question

Table 3.1 lists hypotheses underlying this research.

1. Particular areas of articles may tend to contain nouns closest to the overall subject matter of a text.
2. Repetition of these nouns throughout the article may indicate that a sentence containing these nouns is close to the main subject of a text. Since language of judgement was defined as a subset of language closest to the overall subject matter of a text, presence of these repeated nouns may indicate language of judgement.
3. Rhetorical relations may act as signals for language of judgement.
4. When language of judgement exists within a text (following hypothesis 2), different rhetorical relations may be associated with shifts into and out of judgemental language.
5. Use of language of judgement (and exclusion of other language in a text) may improve the accuracy of a semantic orientation classifier.

Table 3.1: Key research hypotheses

Given these hypotheses, the research question is:

> *Can machine learning features based on repetition of nouns from key areas of a text and on rhetorical structure theory increase the accuracy of a semantic orientation classifier that assesses an opinion piece author's approval or disapproval of the piece's subject matter.*

Three of the hypotheses in table 3.1 map directly onto types of machine learning feature. Underlying these hypotheses is the assumption that, assuming these hypotheses are true, they will produce useful features. The following sections define:

- noun repetition features (hypothesis 2)
- rhetorical relation features (hypotheses 3)
- features combining noun repetition and rhetorical relations (hypothesis 4)

3.2 Noun repetition features

The literature review previously discussed the possibility that sentences containing repetitions of nouns found in an article's lede may tend to be language of judgement. Figure 3.1 shows a simple procedure to create such a feature.

Feature name: SimpleLede

Procedure:

1. Extract nouns and noun phrases from the article lede

2. Mark sentences containing any of those nouns or noun phrases with the feature

Figure 3.1: SimpleLede Feature Definition

The “Simple” part of the “SimpleLede” feature name indicates that the feature is derived from information held within the text that it tags (as opposed to information found elsewhere in the text, as is the case for tags dealing with shifts into or out of judgemental language).

This feature, along with other features used to classify language as potentially judgemental or non-judgemental, is passed in a vector describing each sentence to a machine learning algorithm. This vector holds numeric (real number) values indicating the number of times that each feature is present in the sentence. The number of times each feature is present is counted in an attempt to pass information about the strength of participation in the subject matter of the article.

It is noted that this and all other features defined in this research operate on sentence level units. Section 4.2.4 considers the choice of this unit size.

Article ledes are more complicated than previously described in the literature review. As Potter (2006, p. 25) notes:

> "There are two basic types of leads: hard and soft. A hard lead summarizes the essential facts of the story... while a soft lead may set the scene or introduce a character"

The lede shown above is a hard lede. An example of a soft lede is found in the previously discussed article about Sienna Miller:

> '“BRING out your dead!” was the cry of those unfortunate men whose job it was to collect each night the victims of the Great Plague of 1665 before driving the carts holding the piles of bodies to the mass graves which would be their final resting place.' (Burchill, 2008)

The writer goes on to explain:

> '“Bring out your dead relationships/marriages/ whatever!” could be the cry whenever Sienna Miller walks by"

and so moves away from metaphorical description towards the main subject of the article.

When a soft lede is used, an article's subject may not be raised immediately. Given this, figure 3.2 describes another feature type based on extraction of nouns from the paragraph following the lede (once again, this will quite often only be a single sentence).

Feature name: SimplePostLede

Procedure:

1. Extract nouns and noun phrases from the paragraph following the article lede

2. Mark sentences containing any of those nouns or noun phrases with the feature

Figure 3.2: SimplePostLede Feature Definition

This seems appropriate for the above example since the nouns "relationships", "marriages" and "Sienna Miller" are found after the initial sentence.

There is evidence from other work on semantic orientation that a semantic feature's location within a text may usefully affect its weight. Taboada and Grieve (2004) find placing greater weight on features located two-thirds of the way through a text increases the accuracy of a sentiment classifier. Further reflection on article structure allows definition of two additional features for accessing potential subject matter language.

Feature name: SimpleTitle

Procedure:

1. Extract nouns and noun phrases from the article title

2. Mark sentences containing any of those nouns or noun phrases with the feature

Figure 3.3: SimpleTitle Feature Definition

The feature defined in figure 3.3 seems justifiable as an article title is unlikely to contain language of digression or counterargument. It is thus likely to contain nouns close to the subject matter of the article.

Feature name: SimpleFinal

Procedure:

1. Extract nouns and noun phrases from the article's final paragraph

2. Mark sentences containing any of those nouns or noun phrases with the feature

Figure 3.4: SimpleFinal Feature Definition

The feature in figure 3.4 has a somewhat weaker justification. While this feature can be justified by arguing that the final paragraph of an article may sum up the subject under discussion, review of corpus texts indicates that articles may often shift away from their main subject by this point into digressive language.

The multiple subjects present in Blunkett's (2008) article were already discussed as an example of high-level incoherence. These multiple subjects also lead to digressive language in the final paragraph(s) of the article.

Considering repetition of language itself gives rise to an additional feature for looking for language of judgement. It is hypothesised that nouns closest to the subject matter of an article are most likely to be repeated, for example an article about a celebrity is likely to include frequent mention of the noun phrase for that celebrity's name. Figure 3.5 describes such a feature.

Feature name: SimpleCommon

Procedure:

1. Extract nouns and noun phrases from the article
2. Count the instances of each noun or noun phrase
3. Retain nouns and noun phrases that occur more than once
4. Mark sentences containing any of those nouns or noun phrases with the feature

Figure 3.5: SimpleCommon Feature Definition

The choice of the number one in the above procedure is to some extent arbitrary. Given the hypothesis that areas of digression and counterargument will often tend to be brief, this value is chosen to de-emphasis very short areas of digression or counterargument.

3.3 Features based on rhetorical relations

These features describe potential for particular rhetorical relations to either:

- hold language of judgement
- start or stop an area of language of judgement

This section describes how rhetorical relations can be obtained for these feature sets and then gives definitions for these two types of feature set.

3.3.1 Discovery of rhetorical relations

As previously mentioned in section 1.5, Marcu and Soricut's (2003) SPADE parser will be used to discover rhetorical relations within each sentence. For a given parse, each

sentence is classified according to the highest level nucleus/satellite (or nucleus/nucleus) pair returned by SPADE as this highest level relation is assumed to be most important. This choice is in line with the work of Taboada and Voll (2007).

For a rhetorically simple sentence such as "Triple H may have married into the 24/7 life of a McMahon Family member, but he likes to go home every now and then" (Heyman, 2008), SPADE returns the output shown in figure 3.6. This is easily classified as a Contrast relation.

```
(Root (span 1 2)
  ( Nucleus (leaf 1) (rel2par Contrast)
(text _!Triple H may have married into the 24/7 life of a McMahon Family member ,_!))
  ( Nucleus (leaf 2) (rel2par Contrast)
(text _!but he likes to go home every now and then ._!)))
```

Figure 3.6: SPADE output for a simple rhetorical relation

Figure 3.7 shows a parse by SPADE of a more complex sentence, the previously discussed lede from Burchill (2008).

```
(Root (span 1 5)

  ( Nucleus (leaf 1) (rel2par span)

(text _!“BRING out your dead ! ” was the cry of those unfortunate men_!) )

  ( Satellite (span 2 5) (rel2par Elaboration)

    ( Nucleus (leaf 2) (rel2par span)

(text _!whose job it was to collect each night_!) )

    ( Satellite (span 3 5) (rel2par Enablement)

      ( Nucleus (leaf 3) (rel2par span)

(text _!the victims of the Great Plague of 1665 before driving the carts_!) )

      ( Satellite (span 4 5) (rel2par Elaboration)

        ( Nucleus (leaf 4) (rel2par span)

(text _!holding the piles of bodies to the mass graves_!) )

        ( Satellite (leaf 5) (rel2par Elaboration)

(text _!which would be their final resting place ._!) )

      ))))
```

Figure 3.7: SPADE output for nested rhetorical relations

Given the decision to use the highest level relation in the sentence, this is classified as an ELABORATION relation (with span 1 forming the nucleus of this relation and spans 2-5 the satellite).

A limitation of SPADE is that its RST trees only cover individual sentences. Use of discourse markers, linking words such as “but” and “although”, provides a means to

extend coverage of relations over sentential boundaries. Taboada (2006) considered the frequency that relations are signalled by discourse markers for CONCESSION relations in the RST corpus, a set of articles taken from the Wall Street Journal described in Carlson et al. (2002) and found that CONCESSION relations are likely to be signalled (90.35%). The same article provides a set of markers used for detection of CONCESSION relations: these markers are also used to tag sentences.

3.3.2 RST features seeking language of judgement

These features explore potential association between judgemental language and rhetorical relations. Figure 3.8 gives an example feature that tags CONTRAST relations.

Feature name: SimpleContrast

Procedure:

1. Mark sentences containing CONTRAST relations with this feature

Figure 3.8: SimpleContrast Feature Definition

3.3.3 RST features seeking shifts in language of judgement

These features describe the potential for a shift into or out of potential language of judgement. Two sub-families of algorithms are defined: each family contains individual algorithms for each possible RST feature type (one type for each kind of relation returned by SPADE and another type for all language classified as concessive using the discourse markers). The first sub-family targets shifts out of language of judgement and the second targets shifts into language of judgement.

All algorithms are run after an article has been tagged with the first and second sets of judgement-sensing features (for example SimpleTitle and SimpleContrast). Each algorithm is run once for each tag belonging to the first set of features found in the marked-up article.

For example the BetweenTitleAndContrast_x feature is defined to explore the possibility that a SimpleContrast tag may mark the end of a piece of potentially judgemental language started by a SimpleTitle feature. This is the behaviour required for example 1 (the article 'Why Vince McMahon loves John Cena' (Heyman, 2008)) given in section 1.5.

In practice, the procedure in figure 3.9 would be executed for every sentence that has already been tagged SimpleTitle.

FeatureName: BetweenTitleAndContrast_x

Initial State:

An array exists containing an element for each sentence in the document. Each array element holds a (possibly empty) set of tags. Before the procedure executes all "Simple" semantic and rhetorical features have been tagged (other features indicating potential shifts into or out of language of judgement may have been tagged as well). A SimpleTitle feature has been located which will be used as a starting point in the document.

Procedure:

Initial Setup

1. Save the current state of all sets of tags in the array to allow the algorithm to revert to the initial state if necessary.

2. Create a variable *distance* with the value 0. This variable is used to track the distance (measured as a count of the number of sentences) from the initial sentence containing the starting-point SimpleTitle feature.

3. Create a variable *featureStrength* and set this variable with the numeric value associated with the SimpleTitle feature.

4. Create a variable *sentenceNumber* equal to the index number of the tag set containing the SimpleTitle feature of interest in the tag sets array. This variable will track the current sentence under inspection by the algorithm.

Main Loop

5. Loop over steps 5.1 to 5.3, incrementing the *distance* variable at the end of each iteration of the loop until *distance* exceeds *maxSentences*. The value m*axSentences* is a predefined constant representing the maximum number of sentences to be scanned. If the loop is exited due to *distance* exceeding *maxSentences*, revert the state of all sentence tags to the state saved in step 1 of this procedure and exit the procedure with no changes made.

5.1 Move to the set of tags for the next sentence in the document: increment the *sentenceNumber* variable.

5.2. If the array element indexed by the current value of *sentenceNumber* is tagged with a SimpleContrast feature then exit this procedure. In this case, retain any changes made by the procedure.

5.3.. Create a tag of the form BetweenTitleAndContrast_x (where the value of *distance* replaces the place-holder character 'x') and assign it the numeric value held in *featureStrength*. Add this tag to the set of tags held for the current sentence.

Figure 3.9: BetweenTitleAndContrast Feature Definition

To clarify the operation of the algorithm, before execution the sentences in example 1 have the initial tag-sets:

(S1: {SimpleTitle(1)}, S2: {}, S3: {}, S4: {SimpleContrast})

(for clarity, unrelated tags found in reality are not shown).

The above notation indicates the first sentence has previously been tagged once with the SimpleTitle feature (the number in parentheses indicates the presence of a single

SimpleTitle feature in this sentence) and the fourth sentence has been tagged with the SimpleContrast feature.

After the procedure completes, the final state would be:

(S1: {SimpleTitle(1)}, S2: {BetweenTitleAndContrast_1(1)}, S3: {BetweenTitleAndContrast_2(1)}, S4: {SimpleContrast})

Table 3.2 shows the text from example 1 once more, illustrating initial and final sets of tags after adding BetweenTitleAndContrast_x features.

Article Subject: John Cena *Article Title:* "Why Vince McMahon loves John Cena" *Overall orientation:* Positive			
Sentence number	Initial Tag set	Final Tag set	Text (annotated with language of judgement)
1	{ SimpleTitle(1) }	{ SimpleTitle(1) }	=/Cena is a workhorse.
2	{}	{ BetweenTitleAndContrast_1(1) }	=/He's a tireless promotional machine. And the project, event, DVD, pay per view, film, CD, and merchandise he promotes are all branded "WWE".
3	{}	{ BetweenTitleAndContrast_2(1) }	=/There's not one single wrestler I've met in the past two decades with Cena's drive, ambition and determination to give every fibre of his existence to the company.
4	{ SimpleContrast }	{ SimpleContrast }	Triple H may have married into the 24/7 life of a McMahon Family member, but he likes to go home every now and then

Table 3.2: Example 1 Revisited - 'Why Vince McMahon loves John Cena' (Heyman, 2008)

Figure 3.10 illustrates the more complex case where some sentences have already been marked with a tag of the form BetweenTitleAndConcession_x. In this example, sentences were previously tagged with BetweenTitleAndConcession_x tags from the SimpleTitle tag in S1 and are now tagged as result of the SimpleTitle tag in S3.

Initial State: (S1: {SimpleTitle(1)}, S2: {SimpleTitle(3), BetweenTitleAndConcession_1(1)}, S3:{BetweenTitleAndConcession_2(1)}, S4: {SimpleConcession})
Final State: (S1: {SimpleTitle(1)}, S2: {SimpleTitle(3), BetweenTitleAndConcession_1(1)}, S3:{BetweenTitleAndConcession_2(1), BetweenTitleAndConcession_1(3)}, S4: {SimpleConcession})

Figure 3.10: Tag creation - independence of features with different distances from initial feature

A key point here is that the algorithm allows generation of independent features according to their distance from different initial SimpleTitle features: in the tag set for the third sentence, the first sentence's SimpleTitle(1) feature generates the BetweenTitleAndConcession_2(1) feature and the second sentence's SimpleTitle(3) feature generates the BetweenTitleAndConcession_1(3) feature.

These independent features express the intuition that the greater the distance between a rhetorical relation and an initial feature, the less likely it is that they are connected. This intuition is encoded in the algorithm as follows:

- A copy of the initial sets of tags is saved before the algorithm is run. If the rhetorical relation is not found within *maxSentences* sentences of the starting point, all changes to sentence tags are discarded by reverting to these initial tags. The assumption is that no instance of the relevant rhetorical relation is close enough to the initial feature to be relevant (the constant chosen for *maxSentences* is 4).

- The *distance* variable allows distinction between tags based on distance from the initial potential language of judgement. Depending on the value of *distance*, tags such as BetweenTitleAndContrast_1 or BetweenTitleAndContrast_2 can be created. The intent is to give the machine learning method information on the closeness of the tags to the language of judgement.

An additional reason to limit tagging to within *maxSentences* sentences of the starting point is to limit the number of features provided to the machine learning method since the number of these features scales as a multiple of *maxSentences*. More significantly, without such a limitation an article which:

- has a particular starting semantic feature present

- lacks the relevant rhetorical feature

will end up with every sentence except the sentence containing the starting feature acquiring a tag based on that relation and the starting feature since no condition for removal of such tags would exist.

The original semantic and rhetorical features are retained for the case where no combined feature is tagged since they are considered to have potential value on their own as markers of language of judgement.

The above description uses the separate variables *distance* and *maxSentences* to simplify explanation. However an implementation could use a single decrementing counter.

Both the above examples describe features to investigate potential shifts out of potential language of judgement. Example 2 in section 1.5 requires features to investigate potential shifts into potential language of judgement: Appendix B presents a similar algorithm to tag potential entry into language of judgement signalled by a Contrast relation that scans upwards in the text from an initial feature found through noun-repetition.

While the suffix of a tag indicates distance from potential language of judgement, the numeric value held by a tag indicates potential strength of that language (since that value is derived from the numeric value found in the initial feature).

Chapter 4 Data collection using human annotators

This chapter discusses corpus selection and then describes methodologies for:

- subject and author attitude annotation
- sentence-level annotation

Two main corpora were created for the research, a training corpus used by the machine learning phase and a testing corpus used to obtain final results. Both corpora were looked at by two annotators in an attempt to reduce and quantify annotator bias: a third small corpus of articles was used for initial inter-annotator discussions on methodology.

All articles were taken from the British tabloid "The Sun”. As far as the author of this research is aware, this choice of tabloid articles for assessment of author attitude is novel. A final section of this chapter discusses annotation issues particular to this choice.

4.1 Corpus selection criteria

A single newspaper tabloid was used to avoid possible confusion of the classifier by differences in style between newspapers. It is accepted that results of this research are consequently less general given that they are based on a single source and a limited pool of journalists.

Since language of judgement has been defined as language holding an emotion matching the "overall sense (positive or negative)" of the article, all articles are required to have clear positive or negative orientation for author attitude. Different

procedures are used for the training and testing corpora to manage the possibility that texts might be neutral.

An initial attempt was made to select opinion piece articles as opinion pieces seem more likely to have positive or negative author attitude than news articles that purport to simply report events. Articles were considered opinion pieces if they are located in one of the “Columnists” sections of the Sun newspaper website. However a shortage of opinion pieces lead to differences in selection methodology for the different corpora: while the training corpus was augmented with supplementary articles felt to be clearly positive or negative, only opinion piece articles were used for the testing corpus.

4.2 Annotation procedures

The following sections present annotation procedures with particular reference to the consequences of this shortage of texts.

4.2.1 Pre-annotation clean-up

Articles were obtained from the Sun newspaper's website using the website's “print this article” feature which renders an article in a simple text format. This was copied and pasted into a text editor to create a file containing the article. Only the main bodies of articles were used – sidebars containing supplementary information were excluded when present: the most appropriate method to relate sidebars to the overall rhetorical structure of a text is unclear and sufficient information to determine semantic orientation is likely already present in the main body. Some additional text was also removed when present: images captions, place-holder text for

advertisements, author names, publication dates and a final line inviting readers to comment on the article.

4.2.2 Training set annotation: subject and orientation

Given that some non-opinion piece articles were present in the training set, there was a concern that neutral articles would be a source of noise in training the classifier. When these non-opinion piece articles were selected, articles purporting to be purely factual news stories with no clear author opinions were avoided to reduce this effect. Since this selection was based on a single person's opinion, it was felt that some neutral articles might be chosen: elimination of such articles was an aim of the annotation procedure.

A large number of articles (at least 100) was considered necessary to obtain statistically significant results. Accordingly an initial set of 140 articles was gathered to allow for article elimination. As will be discussed 19 of these articles were eliminated by the first annotator then 121 articles were given to the second annotator. 5 further articles were then eliminated in discussion with the second annotator, leaving a final training corpus of 116 articles.

Although classification results are only dependent on measurement against article orientation, both testing and training sets were annotated for subject given the previously discussed potential for difficulty determining the subject of a text. This allowed observation and understanding of impact of potential disagreements on subject choice between annotators on orientation. Annotation procedure stages are now described.

Stage 1: Pre-annotation discussion

An additional corpus of 12 articles was used for initial discussions between annotators about how article subject should be determined. Given multiple possible ways to express a subject, annotators were guided towards simple subject descriptions with the instruction to:

> "use a single non-ambiguous noun if possible. If this noun would be ambiguous, add an adjective to qualify the noun"

Appendix A contains sample instructions given to annotators. The model described in table 1.1 for categorising types of language was also considered by annotators.

As previously discussed in the literature review, for articles such as Burchill (2008), the subject matter chosen for an article can affect the overall direction of sentiment assigned. Given that opposing sentiment classifications resulted from one annotator considering the subject to be the actress Sienna Miller and the other annotator considering the subject to be society's views about her, additional instruction was given to:

> "chose people or places in preference to abstract concepts"

While this may over-simplify some article subjects and reduce practical usefulness of results, this seems a necessary compromise to handle complex articles that may interpreted on multiple levels.

Stage 2: Subject/Orientation annotation by first annotator

Although the first annotator (the author of this research) attempted to collect texts that were only positive or negative,19 articles were discarded from this initial set when the

first annotator was unable to unambiguously determine a non-neutral semantic orientation (further details on why articles were discarded are given in this chapter's final section). Filtering articles in this way reduced demands on the second annotator's time. The remaining 121 articles were then given to the second annotator.

Stage 3: Subject/Orientation annotation by second annotator

The second annotator was offered a choice of positive, neutral or negative orientations for articles. Articles tagged as neutral by the second annotator were eliminated.

Stage 4: Inter-annotator subject discussion

By this point annotators had chosen article subjects and orientations. Where subject choices were in disagreement, annotators discussed the articles to attempt to resolve this. If annotators could not agree on the subject, the article was discarded. Otherwise the agreed-on subject was used to decide author attitude orientation (for all texts in the training corpus, once subjects were agreed on, annotators did agree on overall orientation).

5 articles were discarded at these stage, leaving 116 articles to be annotated for language of judgement.

4.2.3 Testing set annotation: subject and orientation

40 articles were collected for the testing corpus. Initial text selection aimed to chose equal numbers of positive and negative texts. To avoid introducing bias in article selection, articles were provisionally tagged as positive or negative when they were selected. Once 20 articles were reached of one orientation, only articles of the opposite orientation were selected until its quota of 20 articles was achieved.

While neutral was a permitted orientation choice for the training corpus, annotators were offered a forced choice between positive or negative orientations for the testing corpus. As previously discussed articles were selected from “Columnists” opinion piece sections of the Sun website in an attempt to avoid neutral articles. It was accepted that potential for neutral articles within this corpus may reduce the accuracy of final results. In practice no testing corpus articles were felt to be neutral, however as discussed in section 6.3, annotator assessment of sentiment orientation did vary for a small number of articles due to disagreement on subject.

Both annotators annotated the testing corpus for subject and orientation simultaneously. Texts where annotators did not agree on orientation are not used in calculation of final results: counts of texts where annotators disagree are included as a separate result indicative of the difficulty of determining a single subject for an article.

4.2.4 Training set annotation: language of judgement

Once subject and orientation were agreed on for training set articles, annotators were guided to select language expressing author attitude matching overall text orientation (full annotation instructions for gathering language of judgement are given in Appendix A).

Annotation was performed at sentence level. Clause, sentence and paragraph-level annotation were all considered valid units of analysis but due to time constraints only one annotation unit was chosen. Table 4.1 summarises considerations underlying this choice.

Unit	Advantages	Disadvantages
Clause	Smallest unit possible since the clause is the minimum unit for analysis in RST: most fine-grained coverage	Produces greatest number of units to annotate – limited time available for annotation procedure
Sentence	Alignment with pre-existing work: work on detection of on-topic sentences done by Wiebe et al. (2004) also operates at sentence level. Wiebe et al.'s "on-topic" sentences are considered similar to this work's concept of "Subject matter language (factual or judgemental)"	Loss of resolution compared to clause level annotation: shifts in subject seen in CONTRAST and CONCESSION relations take place inside sentences at the clause level.
Paragraph	Allows of investigation starting from the viewpoint of Potter (2006) that paragraphs may contain a single idea	Many articles composed of single sentence paragraphs: results of paragraph and sentence level analysis identical for these articles

Table 4.1: Considerations for choice of unit of annotation

Post-annotation discussion of sentences with contrasting annotations would be methodologically optimal but was impractical due to time constraints. Given this, possible strategies for dealing with annotator disagreement include:

- Retain language of judgement only if it was annotated by both annotators
- Retain language of judgement if it was annotated by either annotator

While the second strategy would make more language of judgement available, the first strategy gives higher reliability that such language is indeed judgemental. Accordingly this strategy was followed.

4.2.5 Testing set annotation: language of judgement

Language of judgement was also annotated at the sentence level for the testing set using the same procedures as for the training set. While unnecessary for the overall goal of document-level sentiment classification, this allowed investigation of effectiveness of feature sets as discussed in section 6.2.

4.3 Issues encountered during annotation

This section discusses issues related to article subjects that arose during annotation: neutral subjects and ambiguous subjects (affecting orientation or choice of language of judgement). While the issue of neutral subjects only affected the training corpus, the remaining issues arose with both corpora.

4.3.1 Neutral Subjects

As discussed some articles were not chosen (or were discarded during the annotation process) since overall author attitude was considered neutral. Two main reasons were distinguished for this: either the article covered a topic such as a news story in a purportedly factual manner or attitude expressed in the article was not directly attributable to the author.

As an example of the latter, the anonymous article, The Sun Sport (2008) "Gough slams cosy club" contains a criticism of team selection choices for the English cricket

team. However the journalist does not express a point of view directly: instead, as in sentences such as “DARREN GOUGH has slammed England for showing favouritism in their one-day selection policy” attitude is associated with the cricketer Darren Gough. Of five articles discarded during post-annotation discussion, two articles were discarded from the training set for this reason.

Another article, West (2008) 'I am so proud to wear poppy for family's heroes' was excluded as the article was almost completely composed of quotations given to the journalist. Although these quotations expressed a positive attitude towards the subject of the article (wearing a poppy to recognise military sacrifice), this attitude was associated with the interviewee rather than the author. Both annotators considered author attitude to be absent. This also matched annotation instructions for language of judgement to avoid “Statements made by other people than the journalist (t*his includes language inside quotation marks*) ”. Only this one article was discarded for this reason from the training set during inter-annotator discussions (although other articles of this type were avoided during initial article selection).

4.3.2 Subject ambiguity affecting orientation

While disagreement on article subjects leading to disagreement on orientation prompted the instruction to “chose people or places in preference to abstract concepts”, this guidance was not always sufficient. A sub-genre of articles contrasts two individuals, where one individual is portrayed positively and the other negatively. An example of this is Kelly's (2008) 'Obama's barmy to snub Hillary' which praises Hilary Clinton while disparaging Barack Obama. Two such articles were discarded from the training set when annotators could not easily reach agreement on which individual was predominantly the subject of the article.

4.3.3 Subject ambiguity affecting choice of language of judgement

Even when two similar choices of subject lead to the same orientation, annotation of judgemental language may still be sensitive to subject choice. The subject of Ross (2008a) might be the television presenter Noel Edmonds or his television show "Noel's HQ". Accordingly considering the two sentences

> "The maddest TV show you never did see"

and

> "But, of course, the real edge-tipper here was Kim Jong Noel himself" (an insulting comparison between Noel Edmonds and North Korean leader Kim Jong-Il)

for the first subject the first sentence might be considered digression whereas for the second subject the second sentence might be considered digression.

Consequently agreement on subject definition is necessary before annotating language of judgement. Again the guidance to choose "people or places as opposed to abstract concepts" is relevant here. However to reduce potential for lost language of judgement from forced choice of a single subject, guidance for annotation of language of judgement includes the instruction to include:

> "language making a judgement about a person or topic when that language is making an example to reinforce the overall subject "

Chapter 5 Computer-based data processing methods

On entry to the data processing stage, tabloid articles for training and testing corpora were held in text files tagged as either positive or negative for overall author attitude. Corpus articles were also tagged at the sentence level for the presence of language of judgement.

The following sections describe the two stages of data processing:

- Training a classifier to identify judgemental language
- Overall sentiment classification: filtering texts for judgemental language with the classifier followed by sentiment calculation.

5.1 Training the machine learning classifier

Figure 5.1 gives a data flow diagram of the tool-chain for transformation of the training set texts into a model of language of judgement.

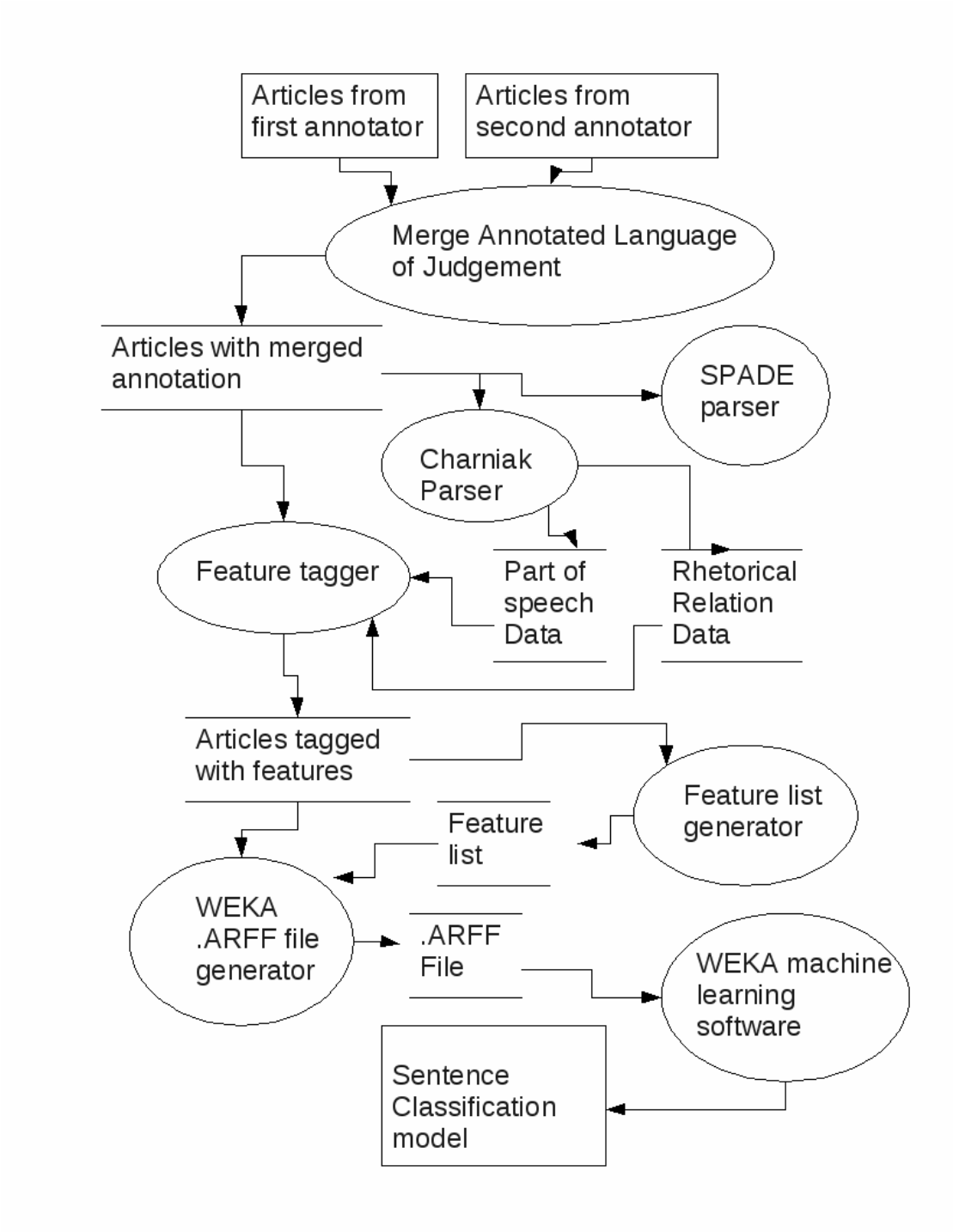

Figure 5.1: Training phase data flow diagram

As discussed in section 4.2.4, articles were assessed for sentences tagged as judgemental by both annotators. The tool-chain generated new versions of the articles with only these sentences tagged as judgemental.

In preparation for tagging with the 3 sets of features described in chapter 3, these sentences were then processed with:

- Charniak's (2000) parser to tag parts of speech. This tags the nouns needed to identify the simple semantic features (and also tags adjectives that will be used in the classification phase).

- Soricut and Marcu's SPADE parser (2003) to identify rhetorical relations.

- Code to check for Taboada's (2006) list of discourse markers for concessive language.

Additional code then tagged each article's sentences with appropriate features.

Once all training set articles have been tagged, a list of unique feature types is created. This list of unique feature types is combined with the set of all features for all articles to create an input file (.ARFF file) used by the WEKA machine learning software (Witten and Frank, 2005). WEKA generates a model of language of judgement from this file.

A number of classifiers are available inside WEKA. While Taboada and Voll(2007) used WEKA's implementation of the Id3 decision tree algorithm, the presence of numeric attributes in the data set prevented its use.

Instead given Pang et al.'s (2002) previously discussed finding that SVMs perform well, albeit in calculating whole document sentiment without a filtering stage, an SVM classifier was used. Cortes and Vapnik (1995) introduced the underlying technology of SVMs by stating:

> "The support-vector network is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high- dimension feature space. In this feature space a linear decision surface is constructed"

Beyond this brief description, internal operations of SVMs are largely out of scope of this study. However tuning of SVMs will be discussed as relevant. Section 6.2.1 and appendix C discuss experiences following the methodology of Hsu et al. (2008) to select a classifier.

It is noted that while other studies use Bayesian classifiers, these were avoided as inappropriate since some features are not independent. This is most clearly the case for features describing potential entry or exit from language of judgement as other features are used in their creation.

An advantage of WEKA is that it provides facilities for K-fold cross-validation which allows some experimentation on training set data: further details are found in section 6.2. WEKA also allows filtering subsets of observations from the input data: these observations can be chosen either randomly or by specific feature type. Testing the ability of a classifier to learn from differently-sized random subsets of features allows generation of learning curves to assess if features do represent patterns in the data

(Russell and Norvig, 1995, p. 538). Filtering by specific feature type facilitates investigation of the relative value of different feature types.

5.2 Document-level Sentiment Classification

This section describes methodologies for generating baseline results using Turney's method followed by results using the machine learning filter.

5.2.1 Baseline results

Each document was processed with the following steps:

1. Charniak's (2000) parser identified adjectives.

2. Calls to the Yahoo Search API obtained hit counts for adjectives in combination with predefined words having positive or negative connotations.

3. Turney's method (discussed in section 2.1.3) was used to calculate semantic orientation for all adjectives in the article.

4. Per-document semantic orientation was obtained by averaging semantic orientation values for all adjectives.

Once average semantic orientations were available for all training set articles, a normalisation factor was chosen by investigating numbers of correct articles for different factor values. This factor is used to assess positive or negative semantic orientation: articles with per-document semantic orientation equal to or greater than this value are considered positive, articles with per-document semantic orientation below this value are considered negative.

Two methods of deriving this factor were explored. Following Taboada and Voll (2007), the factor was chosen by maximising the overall number of correctly classified articles. The current research also aimed to select a factor which both classified a high number of articles correctly and did not excessively penalise positive or negative articles. This fairness was sought by minimising the difference between percentages of correctly classified positive and negative articles.

Articles in the testing set were then processed to obtain per-document average sentiment orientation values. The normalisation factor obtained in the training phase was applied to these values and articles classified as positive or negative. Overall classifier accuracy was then assessed.

5.2.2 Document classification based on language of judgement

Figure 5.2 gives a data flow diagram describing classification of an article when selecting potential language of judgement prior to sentiment calculation.

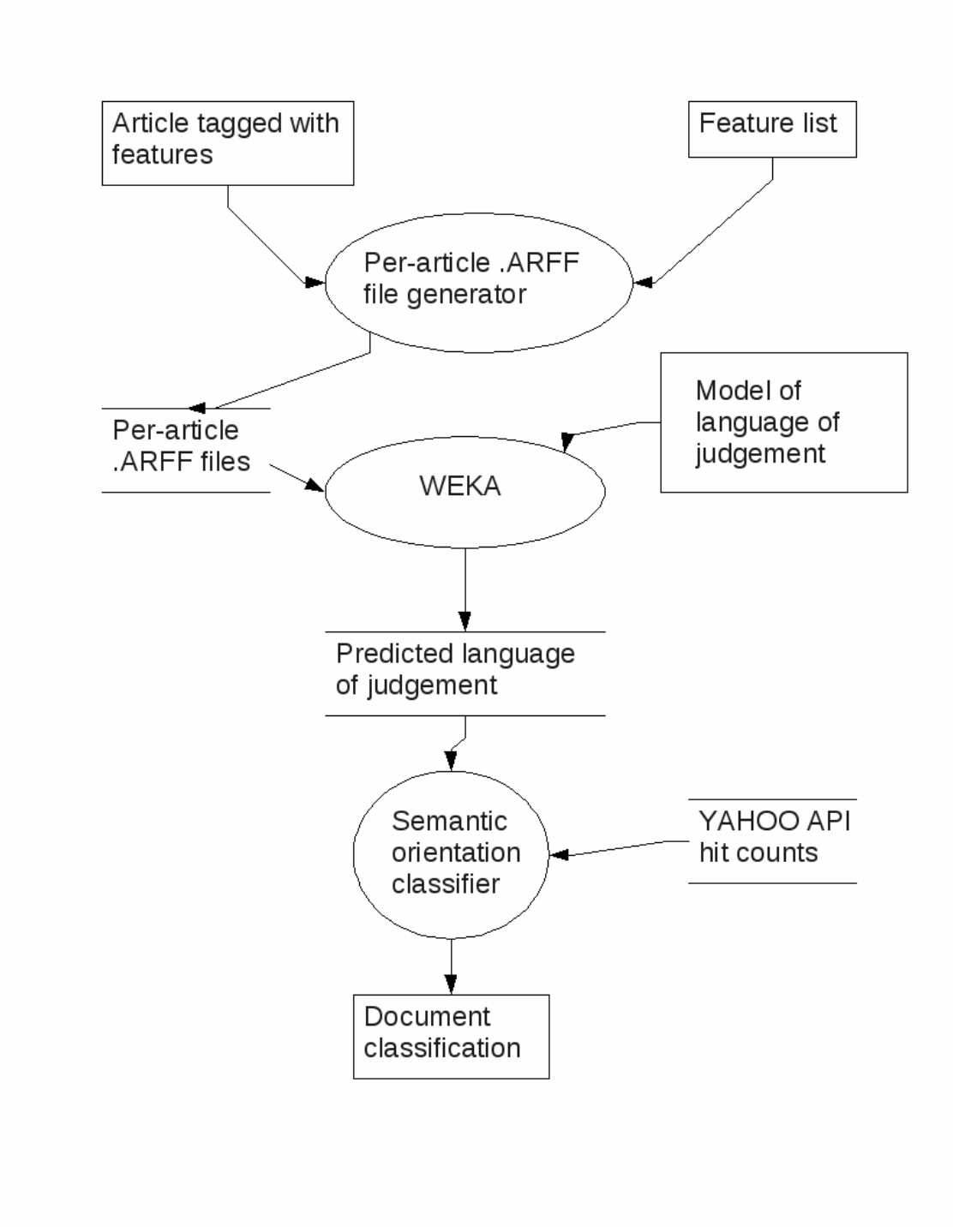

Figure 5.2: Classification phase data flow diagram

Articles in the training corpus were already tagged with features in the learning phase. Tagged articles were combined with the feature list (generated in the training phase) to produce per-article .ARFF files. These files were input to WEKA and the model of

generated in the training phase was used to predict a set of sentences that may be language of judgement. These sentences were then passed through a classifier that assessed per-adjective semantic orientations and calculated overall document sentiment. As for baseline results, a normalisation factor was calculated using the training corpus.

To obtain testing corpus article results, articles were tagged before entry to the classification phase using the same algorithms as in the training set phase. If any features were generated that were not present in the feature list created in the learning phase, these were discarded (an example cause of this would be if the tagger encountered a rhetorical relation not present in the training corpus).

Tagged articles were then passed through WEKA to predict language of judgement. Again, sentiment orientations were assessed using Turney's method for adjectives found in this language and average sentiments calculated. Finally the normalisation factor derived from the training set was used to classify documents. Overall classifier performance was then evaluated.

Chapter 6 Results

The following chapter discusses:

- evidence for the reality of language of judgement.
- usefulness of different feature sets in capturing language of judgement.
- overall ability to perform sentiment classification using language of judgement.

6.1 Evidence for existence of a dimension of judgement

The present research has defined a category of language - “language of judgement” and also outlined how a procedure for how annotators might mark-up such a language. Evidence for appropriateness or otherwise of such a category is provided by the level of agreement between annotators. As Artstein and Poesio (2008) state:

> “Reliability is thus a prerequisite for demonstrating the validity of the coding scheme- that is, to show that the coding scheme captures the “truth” of the phenomenon being studied, in case this matters: If the annotators are not consistent then either some of them are wrong or else the annotation scheme is inappropriate for the data. ”

This work will be used to give a theoretical grounding for discussion of agreement.

The annotation process outlined in Chapter 3 incorporated two tasks:

- overall document annotation for subject and orientation
- sentence-level annotation of judgemental language

Inter-annotator agreement for these tasks is now considered.

6.1.1 Agreement at whole document level

Lack of fixed categories for subject choice causes difficulty assessing agreement given that annotators frequently chose different terms to express most-likely identical subjects – an example pair of annotations seen was “happiness” versus “cheerfulness”. While these annotations were considered to refer to the same concepts in post-annotation discussion for the training set, annotators may have been keen to agree that subject choices were similar and so metrics for subject agreement are not considered meaningful.

Attempting to review testing set subjects may also be vulnerable to similar bias. However, in an attempt at quantification, testing set data was edited such that:

- abbreviated proper names were changed to full names (e.g. Brown became Gordon Brown)
- nicknames were changed to more formal alternatives (e.g. “Hammer's” became the football team “West Ham United”)

After these changes, annotators were found to agree on 60% of subjects. This is considered a minimum value of agreement as many expressions that arguably referred to similar subjects were present.

Since fixed categories exist for orientation, differences in annotator choices are easier to quantify. Numbers of articles discarded during training corpus annotation were previously discussed. However as these articles had already been pre-screened by the first annotator, these numbers do not relate to a randomly chosen corpus and are only meaningful within the context of the annotation process.

Since articles from the testing set did not pass through such screening and were simply chosen by their presence in “Columnist” sections of the website and the desire to have equal initial numbers of positive and negative articles, agreement on orientation may be assessed. However the testing corpus is very small. This is an overall issue for the significance of the results of this dissertation (Taboada and Voll(2007) use a corpus of 400 texts). That said, annotators disagreed on 3 out of 40 articles.

Given the confusion matrix for orientation assessment seen in table 6.1, Cohen's Kappa, k, for inter-annotator agreement is approximately 0.845. Fleisch's Kappa, K, was also calculated (per Artstein and Poesio (2008), this is a more widespread measure of agreement than Cohen's Kappa in computational linguistics) using the individual cases and is again approximately 0.845. This represents good agreement on overall orientation (while Arstein and Poesio indicate meaning of kappa values is subject to debate and sample size remains a concern, this does seem well within the range of good agreement).

		Annotator A		
		Positive	Negative	Total
Annotator B	Positive	15	1	16
	Negative	2	22	24
	Total	17	23	40

Table 6.1: Inter-annotator agreement on text orientation

6.1.2 Agreement on language of judgement

Values of Cohen's and Fleisch's kappa were calculated for inter-annotator agreement on presence or absence of judgemental language. Given the confusion matrices shown in table 6.2 and 6.3, Cohen's Kappa is 0.507 for the testing set and 0.557 for the training set. Fleisch's

kappa was also calculated, with the resulting value of 0.499 for the testing set and 0.554 for the training set.

		Annotator A		
		Judgemental	Non-Judgemental	Total
Annotator B	Judgemental	431	243	674
	Non-Judgemental	81	550	631
	Total	512	793	1305

Table 6.2: Inter-annotator agreement on language of judgement within the testing corpus

		Annotator A		
		Judgemental	Non-Judgemental	Total
Annotator B	Judgemental	1193	543	1736
	Non-Judgemental	257	1655	1912
	Total	1450	2198	3648

Table 6.3: Inter-annotator agreement on language of judgement within the training corpus

These values of kappa indicate moderate inter-annotator agreement, falling well below the 0.8 value suggested by Arstein and Poesio as a good threshold. However these authors do state that they “doubt that a single cutoff point is appropriate for all purposes” and that (in line with the recommendations of others) “Instead, ... researchers should report in detail on the methodology that was followed in collecting the reliability data”.

Given the earlier comment that when annotators disagree, “some of them are wrong or else the annotation scheme is inappropriate for the data”, a tension is noted in the annotation

instructions that may explain the moderate level of inter-annotator agreement. In testing set post-annotation discussion, annotator B stated that annotation results might vary depending on whether she prioritised selecting:

> "language making a judgement about a person or topic when that language is making an example to reinforce the overall subject"

or avoiding:

> "language making a judgement about a person or topic unrelated to the subject of the article".

Annotator B stated that she felt she was tending to select rather than avoid language. Annotator A was not as strongly aware of this tension and the confusion matrices do show that annotator B found more judgemental language (about 31% more for the testing corpus and 20% more for the training corpus) than annotator A. Relatively lower inter-annotator agreement for the testing corpus may also symptomatic of different patterns of annotation between corpora.

As an overall result, since annotators might not agree at all, existence of moderate annotator agreement does indicate existence of language of judgement as defined for this research.

6.2 Ability of features to train a machine learning method in language of judgement

On entry to the machine learning stage, 3644 instances of data were available (Charniak's parser could not parse 4 of the 3648 annotated sentences). 84101 individual features were spread over 1839 separate feature types.

The following sections give an overview of classifier selection using training corpus data followed by evaluation of its performance on both corpora. Relative performance of different sets of features are then considered.

6.2.1 Classifier selection and tuning using the training corpus

This section states main insights from the classifier selection and tuning process (appendix C describes this in depth) and then presents a confusion matrix for the training phase. A key criterion for classifier selection and tuning was that precision is more important than recall, provided recall is not too low.

During the tuning process, the feature creation algorithm was modified to create a single feature where multiple features had previously been created for different values of the distance variable. Values associated with these merged features were added together and assigned to the single combined feature.

Various LIBSVM kernels (Hsu et al., 2008) and the The LIBLINEAR classifier (Fan et al., 2008) were tested with training corpus data. After experimenting with the classifiers' tunable parameters, the LIBLINEAR classifier (with the c parameter set to 0.5) in combination with the merged features was chosen as it was felt to give the best compromise between precision and recall. The precision score for judgemental language was 0.520 with a recall of 0.200. A confusion matrix for 10-fold cross-validation on training corpus data is shown in table 6.4.

		Annotation by classifier	
		Non-judgemental	Judgemental
Annotation by human	Non-judgemental	2230	221
	Judgemental	954	239

Table 6.4: Confusion Matrix for acquiring language of judgement using 10-fold validation of training corpus data

6.2.2 Evaluation of Classifier performance using the testing corpus

Testing corpus results are of greater interest than training corpus results. As Russell and Norvig (p. 538) state:

> "In theory, every time you make a change to the algorithm, you should get a new set of examples to work from. In practice, this is too difficult, so people continue to run experiments on tainted sets of examples."

Consequently a particular concern is that feature generation algorithms were developed using training corpus data. In contrast, all results based on testing corpus data were generated after classifier selection was complete. Consequently testing corpus results possess a rigour that training corpus results do not.

Per the confusion matrix in table 6.5, precision for acquiring judgemental language was 0.405 with recall of 0.162. Although a large amount of judgemental language is missed by this model, a greater proportion of judgemental language is obtained than if machine learning was not used and all language was chosen (judgemental language forms 32.8% of the testing corpus). Consequently useful learning does seem to have taken place. The following sections further consider evidence for this.

		Annotation by classifier	
		Non-judgemental	Judgemental
Annotation by human	Non-judgemental	770	103
	Judgemental	361	70

Table 6.5: Confusion Matrix for acquiring language of judgement from testing corpus data

These scores for precision and recall are however weaker than for the training corpus (where precision of 0.520 and recall of 0.200 were obtained). While tainting of results may be a factor here, potential differences in corpus composition are another possible cause. As discussed in section 4.1, procedures to create training and testing corpora differ: in particular sentimental non-opinion pieces are present in the training corpus. Additionally the small size of the testing corpus might cause texts with unusual rhetorical characteristics to skew results. A final possible cause is the relatively weaker level of inter-annotator agreement on language of judgement for the testing corpus seen in section 6.1.

As testing set recall and precision are lower than for the training corpus, the following sections make use of training set data where results from the testing corpus are considered too weak for conclusions to be drawn. The possible causes stated above are considered to be justification for this. However it is stressed that these results are not seen as proofs, since use of testing set data is methodologically required for this. Where training set results are given, relevant testing set results are also shown.

6.2.3 Evidence for learning of judgemental language

As indicated in section 5.1, learning curves were constructed to assess if learning really was finding patterns. As judgemental language is of more interest than non-judgemental language,

true and false positive rates for acquiring judgemental language are plotted in addition to overall percentages of language acquired.

To create learning curves using testing set data, since cross-validation is not possible, the procedure in figure 6.1 was followed. Creation of learning curves using training set data used 10-fold cross-validation.

For each member of the set of percentage values that will be plotted on the learning curve's x-axis

1 Select 100 random samples of the relevant percentage of the training set

1.1 Train separate models with each sample

1.2 Process the testing set with each model – obtain value of interest for the y-axis

2. Calculate average y-axis value of interest over all 100 samples

Figure 6.1: Procedure for creating learning curves with testing set data

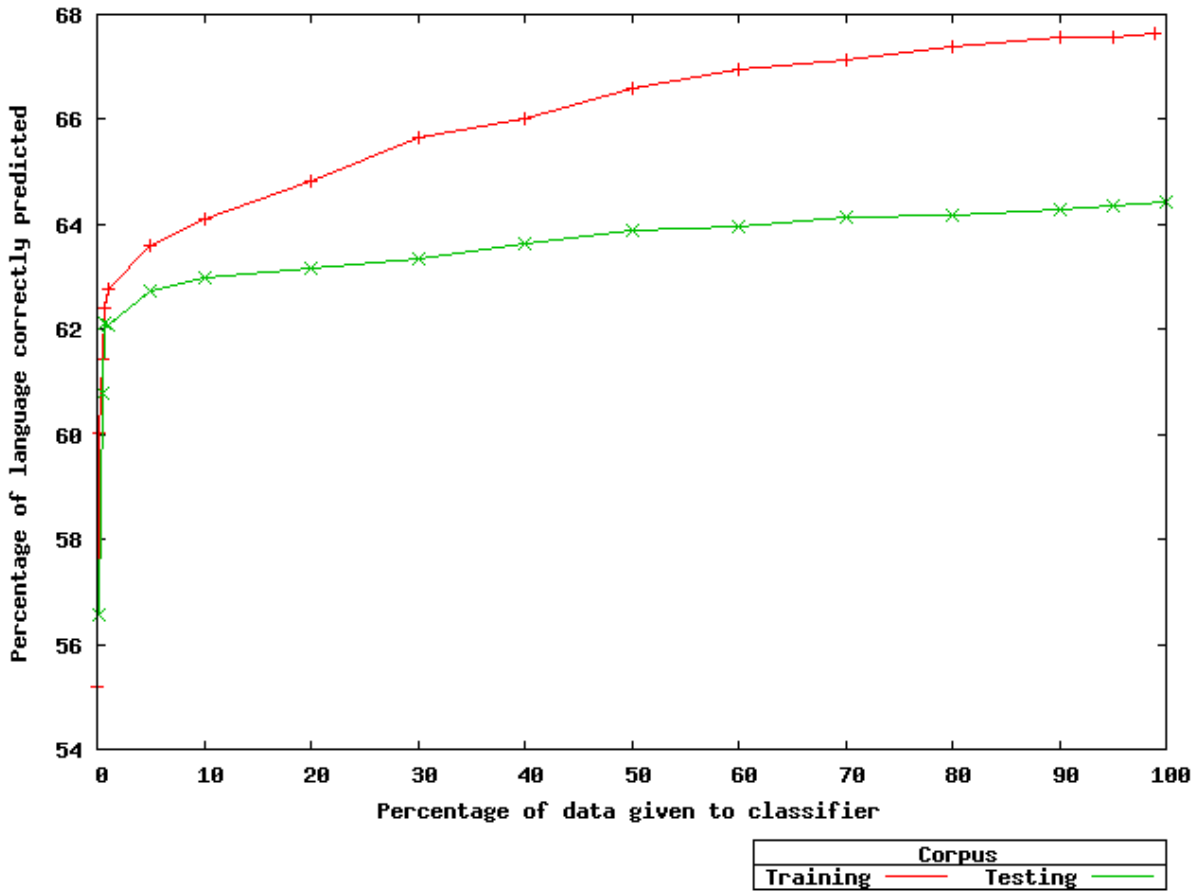

Figure 6.2: Learning curve (percentage correct) for classifying language of judgement

The rising curves for both testing and training corpora shown in figure 6.2 indicate learning does occur for the combination of judgemental and non-judgemental language. However, even for the training corpus (which exhibits relatively better performance), the maximum percentage of correct results obtained using training set data (67.65% with a 99%:1% training:testing split of this data) is only slightly higher than would be achieved by categorisation of all language into the majority, non-judgemental, class (67.26%).

In support of this observation, results were created for 10-fold cross validation of the whole training set using LIBLINEAR and WEKA's ZeroR classifier. ZeroR assumes all instances belong to the most common category. LIBLINEAR classified 66.26% instances correctly and ZeroR naturally classified 67.26% correctly. WEKA's paired T-tester was then run on these

results and no significant statistical difference was found. By this measure the classifier is not particularly powerful.

Figure 6.3 focuses on judgemental language, showing true and false positive rates. The true positive rate indicates the proportion of language considered judgemental by human annotators correctly identified by the classifier (the false positive rate identifies non-judgemental language incorrectly classified as judgemental).

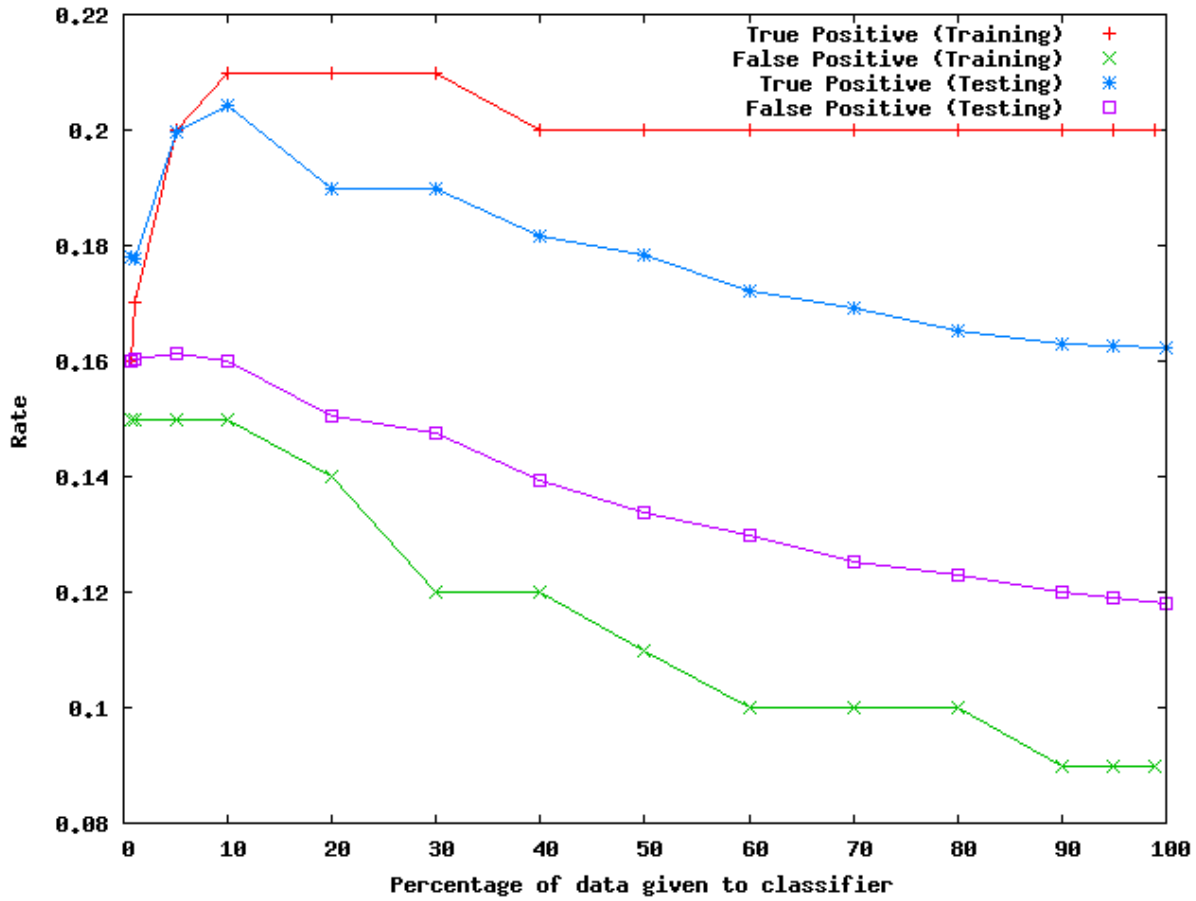

Figure 6.3: True and false positive rates for acquiring language of judgement with testing and training corpora

For the testing set curves, the false positive curve shows evidence of learning but this is not seen for the true positive curve. Falling back to training set data, both curves do indicate that

learning of judgemental language is taking place (albeit to a low maximum true positive rate, consistent with the low recall previously seen).

6.2.4 Performance of separate feature sets: introduction

This section investigates the individual performance of the three feature sets used in the research. As discussed in section 3.1, the research question posed by this project depends on a number of hypotheses. Restating three of these briefly:

1. Repetition of nouns presumed close to the subject matter of the text may indicate language of judgement

2. Rhetorical relations may act as signals for language of judgement.

3. Different rhetorical relations may be associated with shifts into and out of language of judgement.

Since these hypotheses were used to create the three different feature sets used in the research, success by an individual feature set in acquiring language of judgement is evidence in support of the relevant hypothesis. It is stressed that success, particularly where this success is found using training corpus data, is evidence rather than proof - there might be other underlying causes for success of a feature set (analysis of this possibility is out of scope of the present research).

An initial hypothesis underlying the research question is that “particular areas of articles may tend to contain nouns closest to the overall subject matter of a text”. While this is not explicitly explored, evidence found for the first and third hypotheses would act as evidence for this hypothesis as well, given these hypotheses depend on this initial hypothesis.

To evaluate the different features sets, the following sections:

1. Introduce the performance of separate feature sets by discussing their precision and recall for acquiring language of judgement.

2. Consider the effects of removing of individual feature sets from the model produced with all features (since learning is considered to take place when all feature sets are combined).

3. Describe learning curves for the different feature sets.

4. Summarise strength of evidence provided by the feature sets for the different hypotheses.

For brevity, in the following analysis, the different feature sets are referred to as *set 1* (noun repetition), *set 2* (rhetorical relations) and *set 3* (language shifts). Given the decision during classifier selection to use the merged version of the *set 3* features, the following analysis focuses on this variant of the feature set (however data on non-merged features is also discussed).

A limitation of the comparison methodology is that that, unless otherwise stated, testing was done with the LIBLINEAR classifier, using the c=0.5 parameter chosen in classifier selection. Other classifiers (and associated parameters) might be optimal for subsets of the combined feature set.

	Number of Features	
	Training corpus	Testing corpus
Set 1	7465	2704
Set 2	3085	1139
Set 3	73565	26525
Set 3 (merged)	42344	1462

Table 6.6: Feature counts per set (training and testing corpora)

Additionally as shown in table 6.6 , there are large differences in the size of each set (set 2 contains somewhat less features than the total number of instances as SPADE did not always identify a rhetorical relation). Nonetheless full feature sets are used for comparison as opposed to equal sized samples of the different feature sets: While smaller set sizes may have reduced ability to learn, if a feature set has a greater ability to generate features, that ability is considered an intrinsic strength of that set.

6.2.5 Precision and recall for individual feature sets

Table 6.7 shows maximum precision and recall in finding judgemental language achieved for individual features sets.

	Training corpus		Testing corpus	
Feature Set	Precision	Recall	Precision	Recall
All	0.520	0.200	0.405	0.162
Set 1	0.357	0.004	0.333	0.005
Set 2	0.368	0.035	0.463	0.044
Set 3 (pre-merge)	0.423	0.201	0.382	0.434
Set 3 (merged)	0.540	0.148	0.339	0.093

Table 6.7: Precision and recall for acquisition of judgemental language by individual feature sets

Considering testing set data first, set 1 features result in a model with very low recall and poor precision. Set 2 features are relatively best performing with the highest precision (though recall is still very low). Set 3 features show poor precision. None of these feature sets perform well against the testing set data. While the pre-merged set 3 features show much higher recall, use of these features would be tainted as the classifier training phase previously indicated better results through use of the merged features.

Again falling back to training set data, recall for set 1 and set 2 features is extremely low. Set 3 features had better recall. As with the testing set, the pre-merged set 3 features showed the highest recall of the individual sets (but still did not perform as well as the combination of all features). Neither set 1 or set 2 features exhibit higher precision than that achieved by all feature sets combined. The merged set 3 features have higher precision than the combined feature sets but lower recall.

Since testing set data does not provide strong evidence, if training set data is considered of value (given the possible differences between testing and training corpora discussed in section 6.2.2),, set 3 features seem likely to make the greatest contribution to the combined-feature

model. Set 1 and set 2 features seem likely to make a lower contribution, given their very low recall scores.

6.2.6 Effect of removal of feature sets on precision and recall

Table 6.8 shows effects of removing individual feature sets from the combination of all sets (this combined model uses the merged set 3 features).

		Training corpus		Testing corpus	
Feature set removed	Remaining feature sets	Precision	Recall	Precision	Recall
None	All	0.520	0.200	0.405	0.162
Set 1	Set 2 + Set 3	0.523	0.198	0.393	0.148
Set 2	Set 1 + Set 3	0.536	0.156	0.366	0.104
Set 3	Set 1 + Set 2	0.364	0.044	0.542	0.070

Table 6.8: Effects of feature set removal on precision and recall

Considering results obtained using the testing corpus, removing set 1 slightly reduces precision and recall. Removing set 2 reduces precision and recall by greater proportions than occur with the removal of set 1. Removing set 3 features actually increases precision of the model but recall becomes very low – this change in precision may not be meaningful due to the small number of instances classified as judgemental. This testing corpus data provides some evidence for the value of set 2 features in acquiring language of judgement. The small amount of change when set 1 features are removed provides little evidence for their value while the increase in precision with the removal of set 3 features is evidence against their value in determining language of judgement (however poor recall prevents a strong conclusion being drawn here).

Giving the lack of strong conclusions from testing corpus data, results obtained from work with the training corpus are again considered. Here removing sets 1 and 2 slightly increases precision of the model. Removing set 1 slightly reduces recall while removing set 2 reduces recall by a larger amount. Removing set 3 features greatly reduces both precision and recall. Set 1 features are not seen to make a large contribution. Both set 2 and set 3 features seem to contribute to recall, with set 3 features making a greater contribution than set 2.

Again if this data obtained from work with the training corpus is considered of value, the dominance of set 3 features gives evidence that rhetorical relation shifts into or out of language of judgement can be learnt by a classifier and thus for the hypothesis underlying that feature set. In the same way the contribution of set 2 to recall provides evidence that set is useful in acquiring judgemental language and supports the hypothesised value of rhetorical features.

6.2.7 Feature set specific learning curves

An alternative perspective on the strength of different feature sets is seen by considering their individual learning curves. From the metrics discussed in section 6.2.3, true and false positive rates are of interest. Given the weakness of the overall model with respect to WEKA's ZeroR classifier, overall percentages of language acquired using the different feature sets are not discussed.

While learning curves were constructed for the classification of testing set data (using a model derived from the training corpus) as seen in figures 6.4 and 6.5, these curves are either relatively flat or, in the set 1 case, fall to extremely low rates when the full training corpus is used (consequent on the very low recall shown in table 6.7). These curves are not considered

to show learning. These results match previous per-set results for testing corpus data, where neither precision or recall indicated strength of the features.

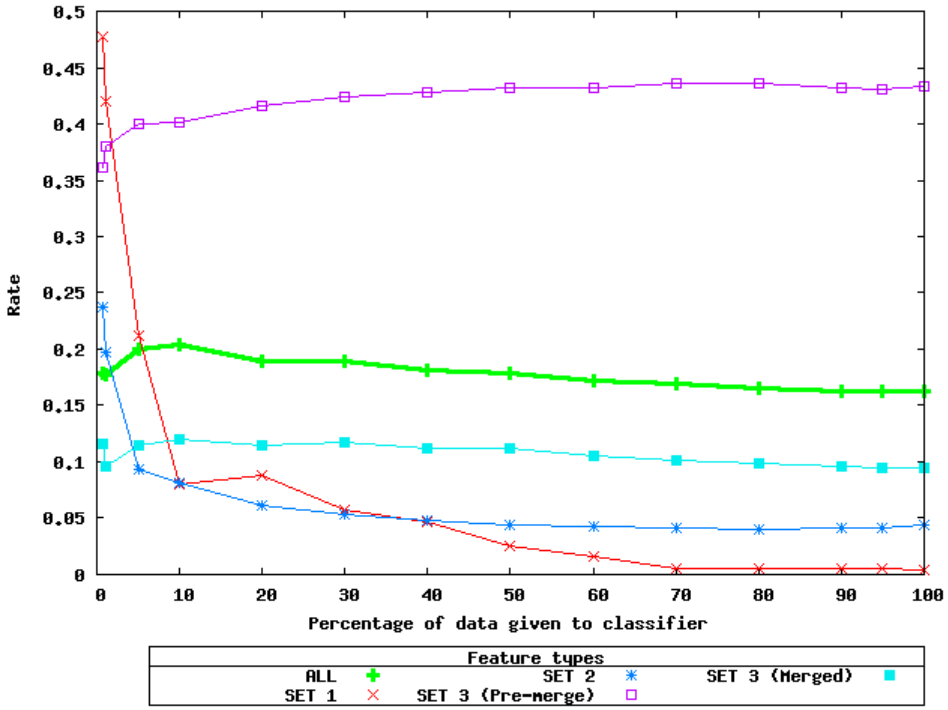

Figure 6.4: True positive rate learning curves for different feature sets assessed using the testing corpus

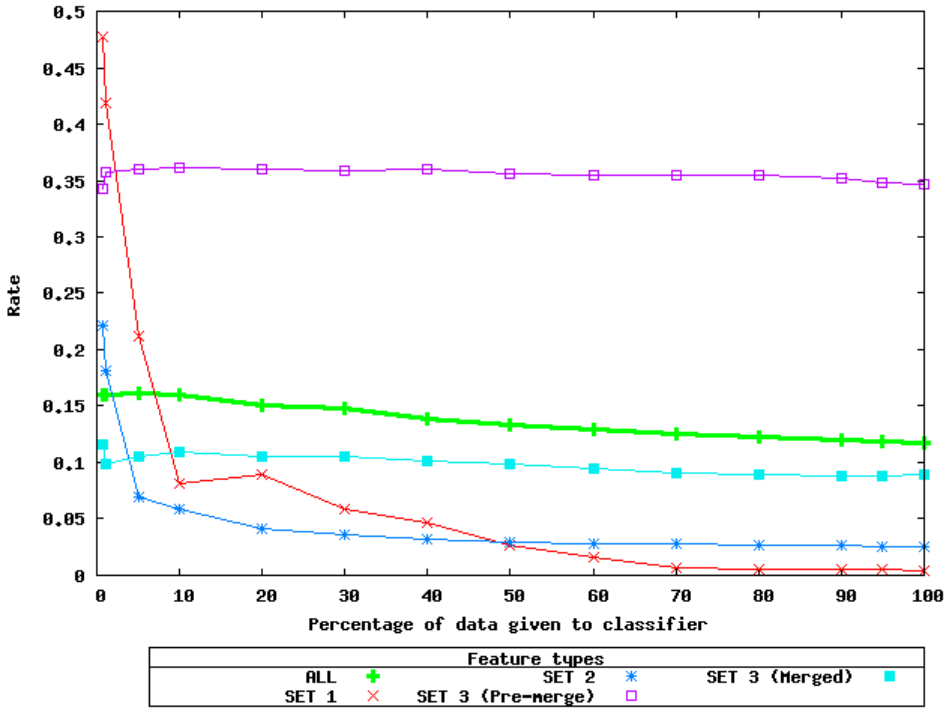

Figure 6.5: False positive rate learning curves for different feature sets assessed using the testing corpus

Accordingly the remainder of this section uses training corpus data only. As in previous sections, use of the training corpus limits the strength of conclusions that can be drawn.

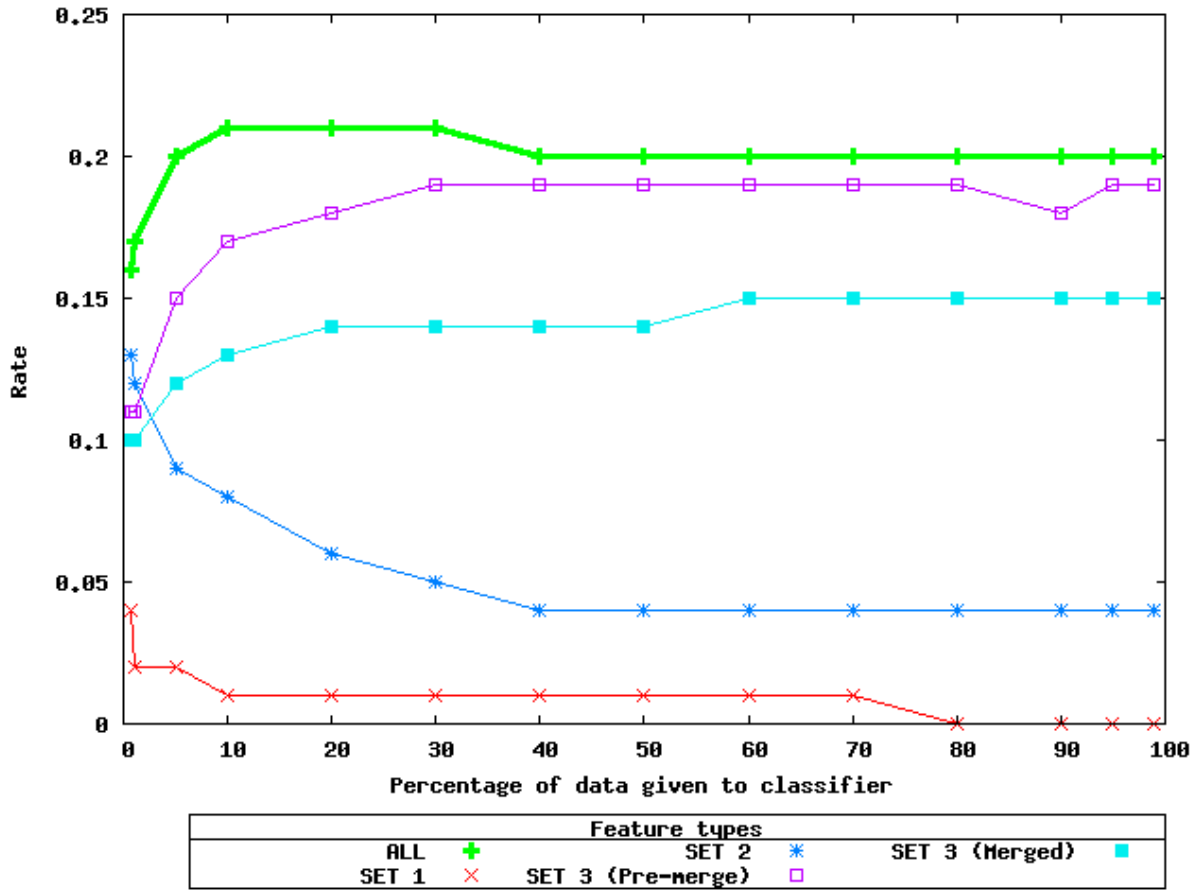

Figure 6.6: True positive rate learning curves for different feature sets assessed using the training corpus

Figure 6.6 shows true positive rate learning curves for the different sets. While the combined features consistently perform best for different percentages of the dataset, learning does occur for set 3 features (more quickly and to a higher rate than for non-merged features). No evidence of learning is seen for set 1 or set 2 features.

Figure 6.7 shows false positive learning curves. Learning, through reduction in the false positive rate, is most apparent for set 2 features. The merged set 3 features also experience learning. The situation is less clear for non-merged features (where the false positive rate initially rises then slowly lowers such that success for training using the entire dataset matches success when 10% of the dataset is used) and set 1 features (the false positive rate is very low but recall for this set is very low and the curve is mostly flat).

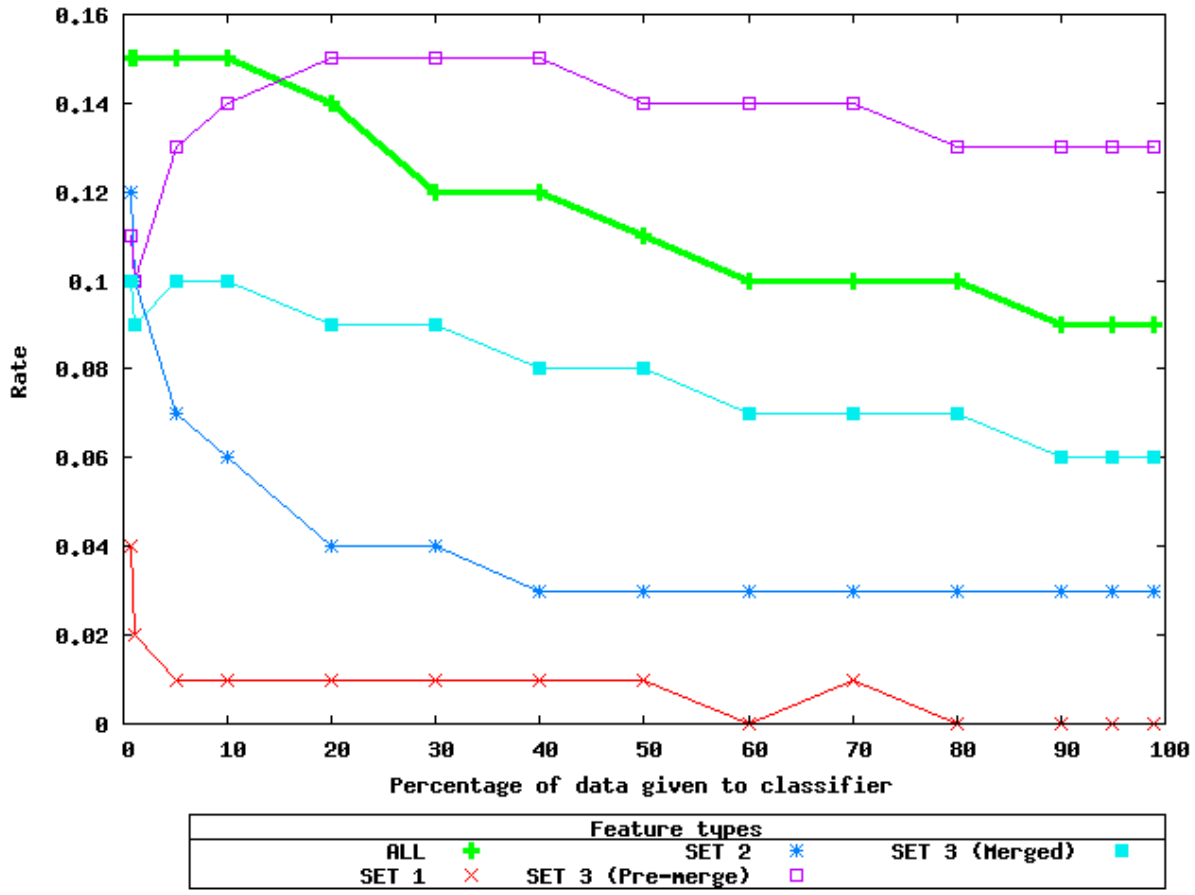

Figure 6.7: False positive rate learning curves for different feature sets assessed using the training corpus

The learning that occurs for the set 3 features (for both true and false positives) is evidence for the hypothesis that shifts in language can be learnt and are of value in classifying language of judgement. The learning that occurs for set 2 features (in the decrease in false negatives) is also evidence for the hypothesis that rhetorical relations can be learnt and are of value.

6.2.8 Degree of evidence for hypotheses from usefulness of feature sets

Evidence for the usefulness of different feature sets was seen for training but not testing corpus data. When using training corpus data, evidence was found for the usefulness of feature sets 2 and 3 and accordingly for the hypotheses that rhetorical relations are of value

acquiring language of judgement, both in themselves and as indicators of shifts into or out of language of judgement. Evidence was not found for the usefulness of the noun-based feature set 1.

6.3 Overall classification

This section discusses results for classification of document-level author attitude. It contains:

- Introductory presentation and analysis of results applying Turney's (2001) method to whole documents.
- Description of optimum possible results achievable assuming perfect knowledge of language of judgement chosen by human annotators.
- Results obtained applying Turney's method to filtered language obtained using the machine learning approach.
- Comparisons between results from whole document and filtering approaches and between these results and the work of Taboada and Voll (2007).

Before presenting results in detail, the composition of the post-annotation testing corpus must be discussed.

As stated in section 4.2.3, an initial selection of 40 texts was made, choosing 20 considered likely to be positive and 20 negative. During annotation three articles initially chosen to make up the positive quota were reclassified as negative (by both annotators independently). Annotators also could not reach agreement on two of the articles chosen to make up the positive quota and one article chosen to make up the negative quota: these articles were discarded. Consequently 37 articles remained after annotation, 22 classified as negative and 15 as positive.

An additional concern is the small testing corpus size. A change in classification of a single article will lead to a 2.7% change in reported results. Given this, results provide evidence rather than statistically valid proof.

6.3.1 Baseline classifier results obtained with Turney's method

As discussed in section 5.2.1, two forms of normalisation factor are used in classifying overall document sentiment, one aiming to maximise the overall number of articles classified correctly (henceforth described as *MAXIMISING*) and the other aiming to classify both positive and negative articles fairly (henceforth described as *BALANCED*). While Taboada and Voll (2007) used a normalisation factor equivalent to this study's *MAXIMISING* factor, they also reported results calculated without a normalisation factor (effectively the same as choosing a zero value for the factor). These results are also reported (henceforth described as *NO_FACTOR*).

Figure 6.8 shows plots of percentages correct (for all articles and articles classified as positive or negative by human annotation) against different factor values using the whole training set. The *BALANCED* factor value of 0.53 is found where curves for positive and negative articles meet. The *MAXIMISING* factor value of 0.43 is found on the highest point of the “all article” curve. Normalisation factors used in the following sections are calculated in the same fashion.

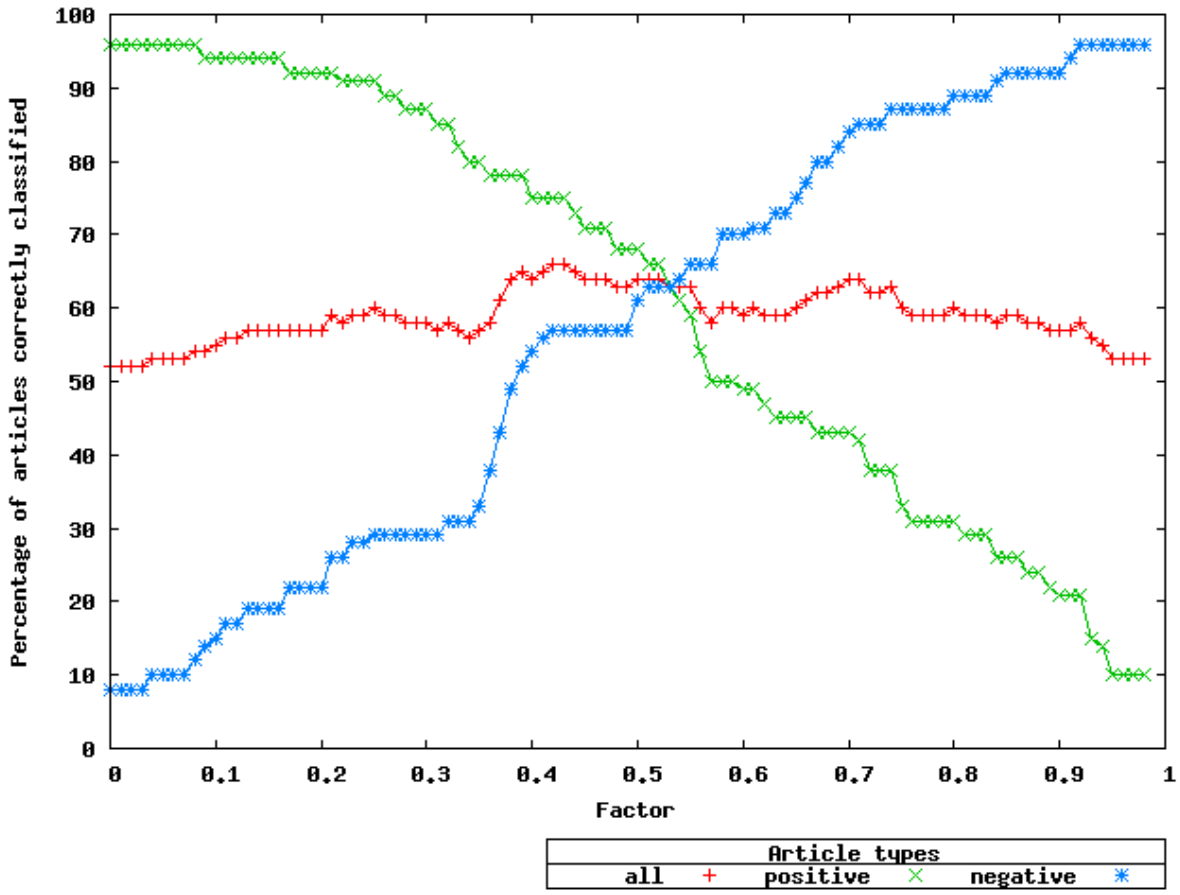

Figure 6.8: Derivation of normalisation factors using training set data

Table 6.9 shows baseline Turney method results for percentages of positive and negative articles as well as for the whole set of articles. Without normalising factors, articles are generally classified as positive, matching the result of Taboada and Voll (2007) and justifying the use of these factors.

		Training corpus: Percentage correct by article category			Testing corpus: Percentage correct by article category		
Factor Type	Factor Value	Positive	Negative	Overall	Positive	Negative	Overall
NO_FACTOR	0	96	8	52	100	0	41
MAXIMISING	0.43	75	57	66	93	38	61
BALANCED	0.53	63	63	63	86	52	66

Table 6.9: Author attitude classification: Baseline results for training and testing corpora

When these factors are applied to testing corpus data, positive articles are still classified with greater success than negative articles. Since the testing set has proportionally more negative articles, overall testing corpus scores may be lower than if the balance between positive and negative articles had been preserved in the annotation stage.

These baseline results are lower than the 72% baseline accuracy achieved with normalisation by Taboada and Voll (2007). However that work also used a hand-ranked dictionary to improve on an initial accuracy of 56% . Consequently direct comparison is not made with these results when considering the effect of incorporating machine learning: instead changes in results with respect to the baseline are considered relevant.

6.3.2 Maximum gains in classifier accuracy assuming perfect ability to acquire language of judgement

To investigate the maximum potential of language of judgement for overall article classification, language considered judgemental by both annotators was extracted from each article and sentiment orientations calculated for adjectives within this language. Per-article

sentiment averages were calculated along with optimal normalisation factors. Percentages of successful article classification were calculated for positive and negative articles as well as the overall collection of articles. Table 6.10 shows classification results.

	Training corpus				Testing corpus			
		Percentage correct by article category				Percentage correct by article category		
Factor	Factor Value	Positive	Negative	Overall	Factor Value	Positive	Negative	Overall
NO_FACTOR	0	87	17	53	0	100	28	58
MAXIMISING	0.47	71	64	68	0.69	73	85	80
BALANCED	0.49	66	67	67	0.59	73	76	75

Table 6.10: Classification results assuming perfect knowledge of language of judgement chosen by human annotators

Compared with results for the training corpus seen in table 6.9, results generated with the *MAXIMISING* factor or without any factor (*NO_FACTOR*) have a smaller gap between percentages of successfully classified positive or negative articles. Improvements in overall classifier accuracy are seen for all three factor choices.

Optimal factor values were recalculated for testing corpus data. The results of applying these factors are shown as an indication of maximum potential results that could be achieved with perfect knowledge of this corpus data.

The use of optimal normalisation factors in the above table may be criticised as contrived. However as figures 6.9 and 6.10 show, this improvement in classifier accuracy holds for nearly all choices of factor.

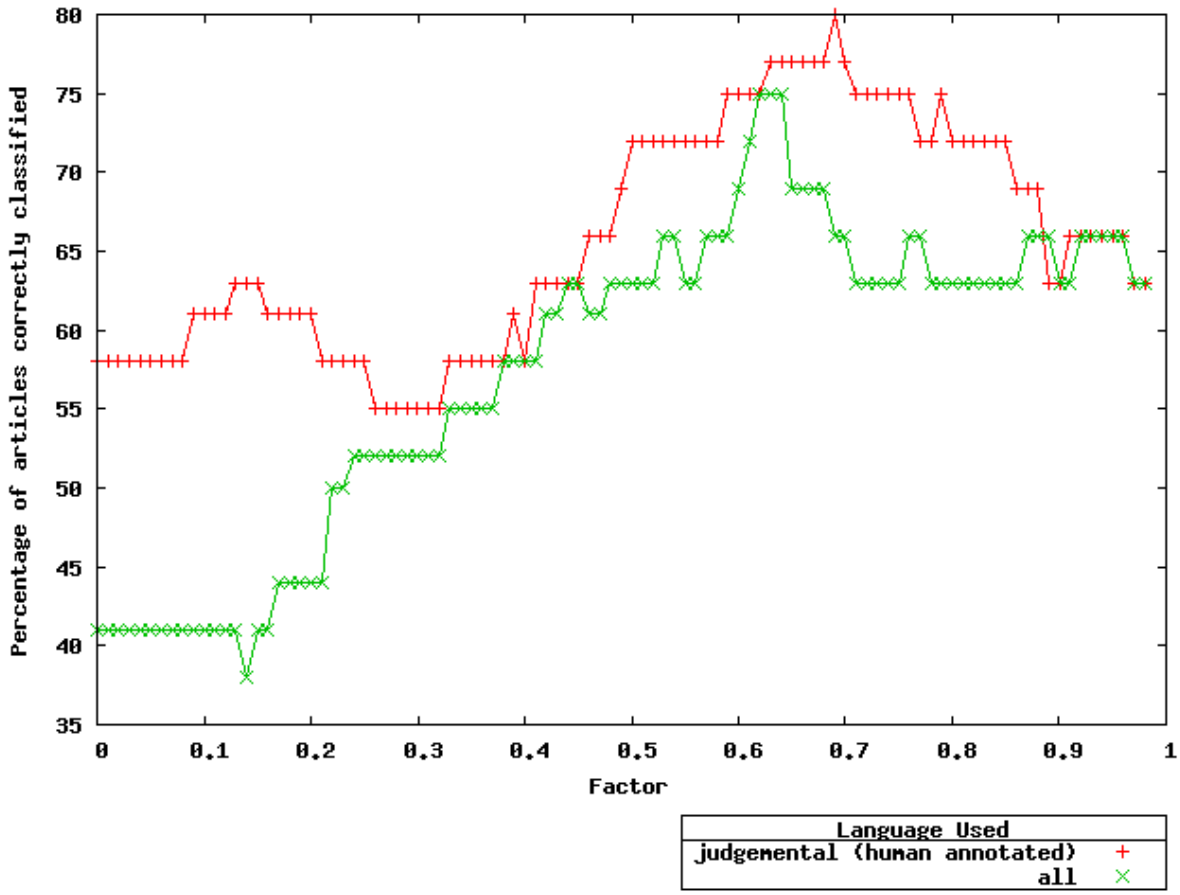

Figure 6.9: Percentages of testing corpus articles classified correctly using human-annotated language of judgement versus all language

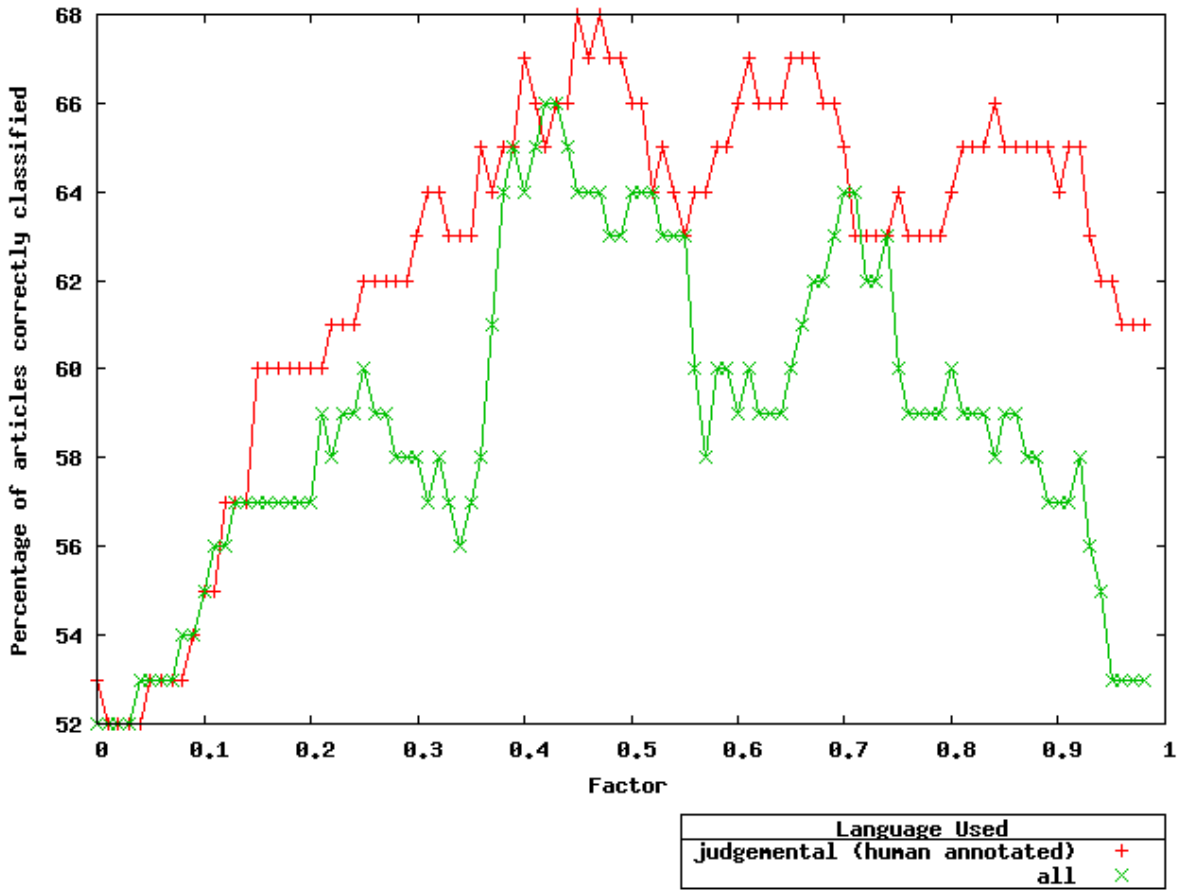

Figure 6.10: Percentages of training corpus articles classified correctly using human-annotated language of judgement versus all language

This general improvement seen over a range of factors when using human-annotated language of judgement is evidence for the hypothesis stated in section 3.1 that:

> "Use of language of judgement (and exclusion of other language in a text) may improve the accuracy of a semantic orientation classifier"

The following section assess the extent that this potential for improvement is realised.

6.3.3 Classifier results using filtered language

Table 6.11 shows percentages of articles classified correctly for training and testing corpora, broken down by human-annotated article classification (positive, negative or overall) and normalisation factor type.

		Training corpus: Percentage correct by article category			Testing corpus: Percentage correct by article category		
Factor Type	Factor Value	Positive	Negative	Overall	Positive	Negative	Overall
NO_FACTOR	0	92	20	55	100	9	47
MAXIMISING	0.51	73	60	66	66	47	55
BALANCED	0.61	61	61	61	60	52	55

Table 6.11: Author attitude classification: Results for training and testing corpora using predicted language of judgement

Some articles were not classifiable as the classifier did not predict any sentimental language for them. This was the case for approximately 7% of the training corpus (9 articles) and approximately 2% of the testing corpus (3 articles). Use of the *BALANCED* factor did preserve fairness in classification of positive and negative articles from the testing corpus (with an 8% difference in relative accuracies compared to a 19% difference for the *MAXIMISING* factor).

Classification of positive articles was more successful than classification of negative articles. As also seen when using Turney's (2001) method alone, overall scores for the testing corpus may be lower than if the balance between positive and negative articles had been preserved in the annotation stage.

6.3.4 Comparative analysis of results

Table 6.12 summarises results for classification of the entire testing corpus using whole (Turney's method alone) and filtered documents (incorporating the machine learning classifier).

	Testing corpus: Percentage of articles correctly classified by approach	
Factor	Whole	Filtered
NO_FACTOR	41	47
MAXIMISING	61	55
BALANCED	66	55

Table 6.12: Author attitude classification: Testing corpus results using whole document and filtered approaches

When no normalisation is used, filtering improves results. However this accuracy of 47% is a poor result and is not considered to be of practical value. Performance is worse for both the *MAXIMISING* and *BALANCED* factors.

Although section 6.3.2 indicated potential for gain from the use of language of judgement in overall article classification, this was not realised in practice. Choice of factor does not seem importance for this reduction in accuracy. As figure 6.11 shows, for testing set data, filtered language almost always performed worse than whole document classification for a range of normalising factors. This may be explained by the results found in section 6.2.2 where prediction of language of judgement within the testing corpus was not seen to achieve high levels of precision or recall.

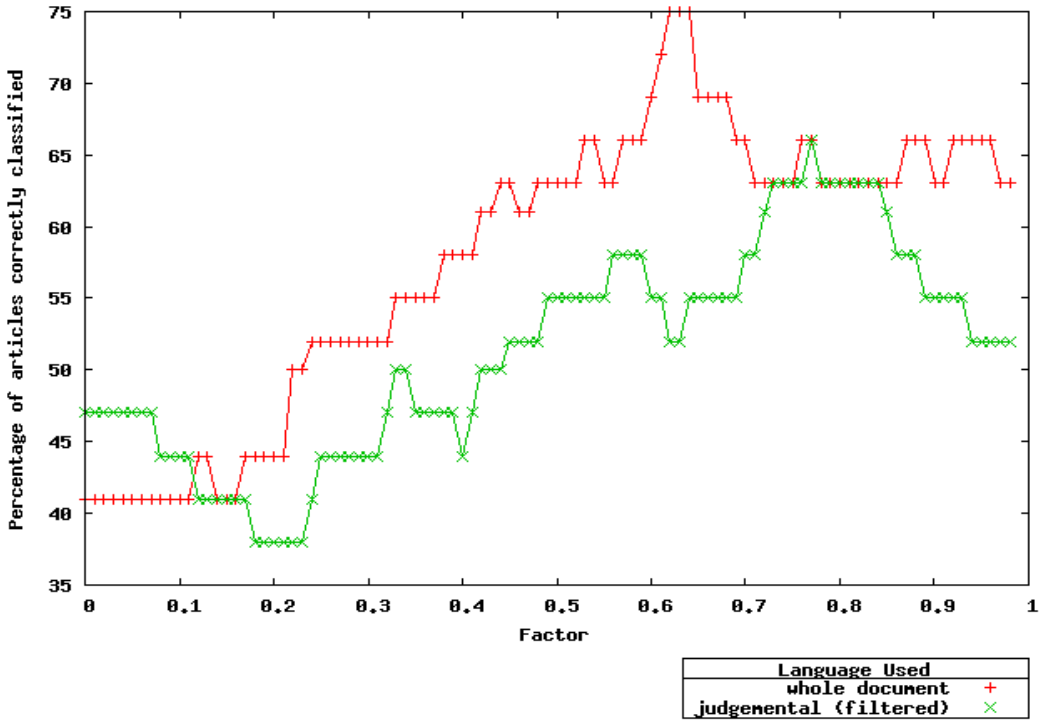

Figure 6.11: Percentages of testing corpus articles classified correctly using whole document and filtered approaches

Given that no relative improvement was gained by filtering language, these results do not show improvements over published work such as Taboada and Voll (2007). As discussed in section 1.4.2, that work employed two filtering methods, one based on rhetorical features and the other based on machine learning of on-topic sentences. The rhetorical feature based approach showed a slight reduction in accuracy: SPADE's 80% accuracy rate was considered a possible cause for this. This is also a source of possible error in the current research. The machine learning approach used by these authors did however lead to improvement in classification accuracy and was thus more successful than the experiment described here.

Chapter 7 Conclusions

Results of the project are reviewed followed by suggestions for further work.

7.1 Project review

This dissertation defined a category of language, language of judgement, with the aim of improving accuracy of author attitude classification. A number of hypotheses, summarised in table 3.1 posited this language was of value and could be acquired through use of semantic and rhetorical features.

The following evidence was found for the hypothesis that focus on judgemental language may improve classifier accuracy (table 3.1, hypothesis 5):

1. High levels of inter-annotator agreement for whole-document classification exist on the specific axis used for measuring semantic orientation used by the project, author attitude (discussed in section 6.1.1).

2. Language of judgement can be annotated on this axis to moderate levels of inter-annotator agreement (measured in section 6.1.2).

3. When human-annotated language of judgement is used to measure per-document author attitude, document classification results are generally better than when using all language in a document (seen in section 6.3.2).

However (as seen in section 6.3.3) improved classifier performance was not achieved in practice. A likely cause of this was low precision and recall for identifying judgemental language in the testing corpus (discussed in section 6.2.2).

Investigating evidence for other hypotheses in table 3.1 gave deeper understanding of the causes of this low recall and precision. It was hypothesised that presence of repeated nouns close to an article's main subject (hypothesis 2) or rhetorical relations (hypothesis 3) might indicate judgemental language. It was also hypothesised that rhetorical relations may indicate shifts into and out of language of judgement (hypothesis 4). However, as discussed in sections 6.2.4-6.2.8, experimentation with testing set data did not find strong evidence for these hypotheses (The first hypothesis in table 3.1, that particular areas of articles may contain nouns closest to the subject matter of a text was not tested).

Experimentation with training set data did however support the third and fourth rhetorical relation-based hypotheses (also in sections 6.2.4-6.2.8). Since training set data was used, this is better seen as indicating value for future exploration of these hypotheses rather than proof of the value of these hypotheses.

Table 7.1 summarises possible causes for differences in ability to acquire language of judgement (taken from sections 6.1.2 and 6.2.2) between testing and training corpora. Potential resolutions possible if more opinion pieces were available are stated. This could be done by collecting articles from multiple newspapers or over a greater time period.

Possible Cause	Potential resolution
Potential tainting of training set results due to use of training set data for feature creation	None: Resolve other causes for poor testing set performance, allowing stronger conclusions to be drawn from the testing set
Presence of non-opinion piece articles in the training corpus: These articles may have different rhetorical characteristics to opinion pieces.	Construct training corpus from opinion pieces
Texts with unusual rhetorical characteristics may be present in testing set and skew results	Construct a larger testing corpus
Relatively lower inter-annotator agreement for the testing corpus may be symptomatic of different patterns of annotation between corpora	A large quantity of opinion pieces would allow an initial combined pool of texts to be divided into training and testing corpora after annotation is complete

Table 7.1: Possible causes of differences in ability to acquire judgemental language between corpora and potential resolutions

Overall moderate levels of inter-annotator agreement (section 6.1.2) and possible incorrect classification of relations by SPADE (section 6.3.4) are additional possible sources of error for results obtained with both corpora.

7.2 Suggestions for future research

Improving the strength of testing corpus results is an obvious direction for future research: this could be done by gathering larger numbers of opinion pieces and implementing the potential resolutions in table 7.1.

While the model presented for categories of language in table 1.1 allows for multiple subjects (as discussed in section 2.1.2) if one subject is considered the main subject and other subjects are considered digression, an annotation scheme that annotated individual subjects separately might better model reality and so improve results. Alternatively performing analysis with a pre-processing stage where articles were separated into component subjects might lead to more accurate results.

As discussed in section 4.2.4, both clause and paragraph level annotation of language of judgement would have been acceptable annotation choices. Annotating language of judgement at these levels may improve results of acquisition of this language.

Finally, given some evidence for rhetorical relations indicating language of judgement and shifts into or out of language of judgement, applying a feature comparison approach to the individual feature types within the associated feature sets may lead to better understanding of the power of these features.

References

Artstein, R. and Poesio, M. (2008) '"Inter-Coder" Agreement for Computational Linguistics', *Computational Linguistics*, volume 34, issue 4 (December, 2008), pp. 555-596.

Battistella, E. (1990) *Markedness: The Evaluative Superstructure of Language*, State University of New York Press, Albany, New York.

Biber, D., Conrad, S. and Reppen, R. (2002), *Corpus Linguistics: investigating language structure and use*, Cambridge, Cambridge University Press.

Blunkett, D. (2008), 'Crying wolf is a risky game', *The Sun newspaper website*, [online], http://www.thesun.co.uk/sol/homepage/news/columnists/blunkett/article1336206.ece [Accessed 18 December 2008].

Burchill, J. (2008) 'Double standard hits Sienna's rep', *The Sun newspaper website*, [online], http://www.thesun.co.uk/sol/homepage/news/article1545432.ece [Accessed 17 December 2008].

Carlson, L., Marcu, D. and Okurowski, M.E. (2002) 'RST Discourse Treebank' [Corpus], Philadelphia , Linguistic Data Consortium.

Cortes, C. and Vapnik, V. (1995) 'Support-Vector Networks', *Machine Learning*, volume 20, no. 3, pp. 273-297.

Crick, A., Haywood, L. and Wheeler, V. (2008) 'World's reaction to Obama's win', *The Sun newspaper website* [online], http://www.thesun.co.uk/sol/homepage/news/the_real_american_idol/article1898567.ece [Accessed 15 December 2008].

Charniak, E. (2000) 'A maximum-entropy-inspired parser', *Proceedings of the NAACL 2000* (April-May), San Francisco, Morgan Kaufmann, pp. 132–139.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin. C.-J. (2008), 'LIBLINEAR: A library for large linear classification', *Journal of Machine Learning Research*, volume 9 (June), pp. 1871-1874.

Heyman, P. (2008) 'Why Vince McMahon loves John Cena', *The Sun newspaper website* [online], http://www.thesun.co.uk/sol/homepage/sport/wrestling/heyman/article1625079.ece [Accessed 29 August 2008].

Hatzivassiloglou, V. and McKeown, K. (1997) 'Predicting the Semantic Orientation of Adjectives', *Proceedings of the 35th Annual Meeting of the ACL* (July), Morristown, Association for Computational Linguistics, pp. 174-181.

Hobbs, J. R, (1985) *On the Coherence and Structure of Discourse*, Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University.

Hsu, C. W., Chang, C. C., Lin, C. J., (2008) *A practical guide to support vector classification* [online], http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf [Accessed 19 January 2009].

Kelly, L. (2008), 'Obama's barmy to snub Hillary', *The Sun newspaper website* [online], http://www.thesun.co.uk/sol/homepage/news/columnists/lorraine_kelly/article1626191.ece [Accessed 11 January 2009].

Kilgarriff, A (2007) 'Googleology is Bad Science', *Computational Linguistics*, volume 33, number 1 (March), pp. 147-151.

Knott, A (1996) *A Data-driven Methodology for Motivating a Set of Coherence Relations* Ph.D. Thesis, Department of Artificial Intelligence, University of Edinburgh.

Lehrer, A. (1974) *Semantic Fields and Lexical Structure*, Amsterdam, North Holland.

Mann, W.C. And Taboada, M. (2005) 'Access to definitive RST documents', *Rhetorical Structure Theory website* [online], http://www.sfu.ca/rst/05bibliographies/access.html [Accessed 7 June 2008].

Mann, W.C. and Thompson, S.A. (1987) 'Rhetorical Structure Theory: A Theory of Text Organization', *Rhetorical Structure Theory website* [online], http://www.sfu.ca/rst/05bibliographies/report.html [Accessed 2 April 2008].

Mann, W.C. and Thompson, S.A. (1988) 'Rhetorical Structure Theory: Toward a functional theory of text organization', *Text*, vol. 8, no. 3, pp. 243-281.

Matveeva, I. and Levow, G. (2007) 'Topic Segmentation with Hybrid Document Indexing', *Proceedings of Empirical Methods in Natural Language Processing (EMNLP) 2007*, pp. 351-359.

Pang, B. and Lee, L. (2005) 'Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales', *Proceedings of the ACL 2005*, Morristown, Association for Computational Linguistics, pp. 115-124.

Pang, B. and Lee, L. (2006) 'Get out the vote: Determining support or opposition from Congressional floor-debate transcripts', *Proceedings of Empirical Methods in Natural Language Processing (EMNLP) 2006*, Morristown, Association for Computational Linguistics, pp. 327-335.

Pang, B., Lee, L. and Vaithyanathan, S. (2002) 'Thumbs up? Sentiment Classification using Machine Learning Techniques', *Proceedings of Empirical Methods in Natural Language Processing (EMNLP) 2002*, Morristown, Association for Computational Linguistics, pp. 79-86.

Picard, R.W. (1995) *Affective Computing* [online], http://affect.media.mit.edu/pdfs/95.picard.pdf [Accessed 25 February 2009].

Potter, D. (2006) *Handbook of independent journalism* [online], http://www.america.gov/media/pdf/books/journalism.pdf [Accessed 26 February 2009].

Ross, A. (2008a) 'Noel reaches for turquoise trackie', *The Sun newspaper website*, [online], http://www.thesun.co.uk/sol/homepage/news/columnists/ally_ross/article1761787.ece [Accessed 11 January 2009].

Ross, A. (2008b) 'Britannia High ... all mingin' all dancin'', *The Sun newspaper website* [online], http://www.thesun.co.uk/sol/homepage/news/columnists/ally_ross/article1903146.ece [Accessed 6 January 2009].

Russell, S. and Norvig, P. (1995) *Artificial Intelligence: A Modern Approach*, New Jersey, Prentice Hall.

Shanahan, F. (2008a) 'Bed-hoppers are screwing us all', *The Sun newspaper website* [online], http://www.thesun.co.uk/sol/homepage/news/columnists/fergus_shanahan/article790620.ece [Accessed 6 March 2009].

Shanahan, F. (2008b) 'Gordon will make JK's £1m vanish', *The Sun newspaper website* [online], http://www.thesun.co.uk/sol/homepage/news/columnists/fergus_shanahan/article1721967.ece [Accessed 18 December 2008].

Shanahan, F. (2008c) 'Bank bailouts and no one's bovvered', *The Sun newspaper website* [online], http://www.thesun.co.uk/sol/homepage/news/columnists/fergus_shanahan/article1748318.ece [Accessed 18 December 2008].

Soricut, R. and Marcu, D. (2003) 'Sentence level discourse parsing using syntactic and lexical information', *Proceedings of Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL)*, Morristown, Association for Computational Linguistics.

Spertus, E. (1997) 'Smokey: Automatic recognition of hostile messages', *Proceedings of Innovative Applications of Artificial Intelligence (IAAI)*, Providence, AAAI Press, pp. 1058-1065.

Taboada. M. and Grieve, J. (2004) 'Analyzing appraisal automatically', *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07)*, Providence, AAAI Press, pp. 158-161.

Taboada, M. (2006), 'Discourse markers as signals (or not) of rhetorical relations', *Journal of Pragmatics*, vol. 38, no. 4 (April), pp. 567-592.

Taboada, M. (2007), *Thumbs Up or Thumbs Down? Detecting Sentiment and Opinion Automatically* [online], http://www.sfu.ca/~mtaboada/docs/Taboada_Sentiment_SFU_2007.pdf [Accessed 7 June 2008].

Taboada, M. and Voll, K. (2007) 'Not all words are created equal: Extracting semantic orientation as a function of adjective relevance', *AI 2007: Advances in Artificial Intelligence*, Berlin, Springer, pp. 337-346.

The Sun Sport (2008) 'Gough slams cosy club', [online], http://www.thesun.co.uk/sol/homepage/sport/cricket/article1935864.ece [Accessed 18 December 2008].

Turney, P. (2001) 'Mining the Web for synonyms: PMI-IR versus LSA on TOEFL'. *Proceedings of the Twelfth European Conference on Machine Learning*, Berlin, Springer-Verlag, pp. 491-502.

Turney, P. (2002) 'Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews', *Proceedings of 40th Meeting of the ACL*, Morristown, Association for Computational Linguistics, pp. 417-424.

Turney, P. and Littman, M.L. (2003), 'Measuring praise and criticism: Inference of semantic orientation from association', *ACM Transactions on Information Systems*, volume 21, issue 4, pp. 315-346.

Turney, P. (2007), *Semantic Orientation - Applications* [online], http://www.apperceptual.com/ml_text_orientation_apps.html [Accessed 7 June 2008].

Wiebe, J., Wilson, T., Bruce, R., Bell, M. and Martin, M. (2004), 'Learning Subjective Language', *Computational Linguistics*, volume 30, issue 3, pp. 277-308.

West, A. (2008) 'I am so proud to wear poppy for family's heroes', *The Sun newspaper website* [online], http://www.thesun.co.uk/sol/homepage/news/campaigns/our_boys/article1899326.ece [Accessed 11 June 2008].

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical machine learning tools and techniques, 2nd Edition*, San Francisco, Morgan Kaufmann.

Wolf, F. and Gibson, E. (2005) 'Representing Discourse Coherence: A Corpus-based Analysis', *Computational Linguistics,* volume 31, issue 2, pp. 249–87.

Yahoo (2009) 'Web Search Documentation for Yahoo! Search', *Yahoo! Developer Network* [online], http://developer.yahoo.com/search/web/V1/webSearch.html, [accessed 7 January 2009].

Index

Appendix A: Training set annotation instructions

Annotation exercise #1 (Text subject and Orientation)

1. Decide if the author is in favour or against the subject.

Read the whole article and decide if the author is in favour or against the main topic of the article.

Articles may discuss multiple topics. If you are in doubt after reading the whole article: Use the title of the article and the first few paragraphs as a guide to decide on the main topic.

If it is not possible to determine this, write that in the section at the top of the article. Otherwise write "in favour" or "against" at the top of the article.

2. Record who or what the article is about ("the subject").

Chose the simplest subject: use a single non-ambiguous noun if possible. If this noun would be ambiguous, add an adjective to qualify the noun.

When choosing the subject, select people or places as opposed to abstract concepts.

Write the subject at the top of the article.

Annotation exercise #2 (Language of Judgement)

1. Consider the subject and author viewpoint provided with the article.

2. For each sentence within the text consider if it contains language which states or implies a judgement about the subject made by the author writing the article.

> If it was decided that the author was in favour of the subject overall, only mark sentences that imply that the author is in favour of the subject.
>
> If it was decided that the author was against the subject overall, only mark sentences that imply that the author is against the subject.

Mark sentences by typing an "=" character at the front of the sentence.

Avoid:

> Language making a judgement about a person or topic unrelated to the subject of the article .
>
> Statements made by other people than the journalist (this includes language inside quotation marks) .

Include:

> Language making a judgement about a person or topic when that language is making an example to reinforce the overall subject.

When in doubt if a sentence is making a judgement or not, include it.

Appendix B: Algorithm to tag features investigating potential shifts into language of judgement

Figure B.1 describes an example feature to explore entry into potential language of judgement, for an area of text between a SimpleContrast feature and a SimpleTitle feature. The description of the algorithm is very similar to the algorithm described in figure 3.9. However in this case the algorithm starts at the SimpleTitle feature and scans upwards looking for a SimpleContrast feature (as opposed to scanning downwards looking for a SimpleContrast feature). Accordingly the only changes are:

- in step 5.1 the *sentenceNumber* variable is decremented rather than incremented
- the name of the feature tag used is different: the tag is titled BetweenContrastAndTitle_x (rather than BetweenTitleAndContrast_x as used for the corresponding tag for potential exit from language of judgement). This is reflected in the feature name and in step 5.3

FeatureName: BetweenContrastAndTitle_x

Initial State:

An array exists containing an element for each sentence in the document. Each array element holds a (possibly empty) set of tags. Before the procedure executes all “Simple” semantic and rhetorical features have been tagged (other features indicating potential shifts into or out of language of judgement may have been tagged as well). A SimpleTitle feature has been located which will be used as a starting point in the document.

Procedure:

Initial Setup

1. Save the current state of all sets of tags in the array to allow the algorithm to revert to the initial state if necessary.

2. Create a variable *distance* with the value 0. This variable is used to track the distance (measured as a count of the number of sentences) from the initial sentence containing the starting-point SimpleTitle feature.

3. Create a variable *featureStrength* and set this variable with the numeric value associated with the SimpleTitle feature.

4. Create a variable *sentenceNumber* equal to the index number of the tag set containing the SimpleTitle feature of interest in the tag sets array. This variable will track the current sentence under inspection by the algorithm.

Main Loop

5. Loop over steps 5.1 to 5.3, incrementing the *distance* variable at the end of each iteration

of the loop until *distance* exceeds *maxSentences*. The value m*axSentences* is a predefined constant representing the maximum number of sentences to be scanned. If the loop is exited due to *distance* exceeding *maxSentences* revert the state of all sentence tags to the state saved in step 1 of this procedure and exit the procedure with no changes made.

5.1 Move to the set of tags for the next sentence in the document: decrement the *sentenceNumber* variable.

5.2. If the array element indexed by the current value of *sentenceNumber* is tagged with a SimpleContrast feature then exit this procedure. In this case, retain any changes made by the procedure.

5.3.. Create a tag of the form BetweenContrastAndTitle_x (where the value of *distance* replaces the place holder character 'x') and assign it the numeric value held in *featureStrength*. Add this tag to the set of tags held for the current sentence.

Figure B.1: BetweenContrastAndTitle Feature Definition

Appendix C: Choosing and tuning a classifier for language of judgement

A criterion for assessing classifier performance must first be defined. Given that the purpose of the classifier in the overall research is to predict potential language of judgement which will be mined for sentiment bearing features, this study does not directly evaluate classifier performance for non-judgemental language. For language of judgement, both precision and recall are important.

If recall is too low, insufficient judgemental language will be found to classify an article (or a classification may not be possible at all if no language is found). At the same time, the greater the precision, the more likely it is that language identified as judgemental actually belongs to this category (to the extent that the training data reflects reality). Given this, the criterion used for comparing classifiers is that precision is more important than recall, provided recall is not too low.

This research follows Hsu et al.'s (2008) methodology for training an SVM. Accordingly values in the training data were scaled to between 0 and 1 by dividing each value by the maximum value found for that feature (and consequently articles classified using the trained model had their relevant features adjusted by these per-feature values before they were input to WEKA). Also following Hsu et al., LIBSVM's RBK kernel (a kernel is the core classification algorithm used by an SVM) was chosen to assess initial performance.

Initially this kernel produced zero recall for language of judgement, classifying all language as non-judgemental. The easy.py script provided by the same authors and packaged with LIBSVM was used to perform an automated grid-search for appropriate values of SVM parameters (prior to using this script .ARFF file data was exported from WEKA into LIBSVM's data format). A possible limitation of this tuning method for this research is that it

does not emphasis accuracy of judgemental language, instead aiming for overall classifier accuracy.

After tuning the RBK kernel scored a precision of 0.53 and a recall of 0.096 (all values for precision and recall obtained in classifier comparison tests were derived using 10-fold cross-validation within WEKA).

This section concludes with a summary of this and other test results: The preceding test will be referred to as *LIBSVM_RBK*. Throughout this section test names are noted in italics. All further results have appropriate tuning applied.

Given the poor recall score from the RBK kernel, 3 strategies for improvement were considered and tests performed:

1) Replacement of RBK kernel with a linear kernel within LIBSVM

Hsu et al. advise this when the number of instances is greatly below number of features. This was tested in test *LIBSVM_LINEAR*. Tuning was done using the grid.pl script provided by these authors.

2) Use of LIBLINEAR

When instance and feature counts are both large, Hsu et al. recommend use of the LIBLINEAR SVM instead of LIBSVM, motivating test *LIBLINEAR*. Results produced by LIBLINEAR were tuned by manually adjusting LIBLINEAR's C parameter (affecting the impact of incorrect predictions in training to the model created by the SVM). Values of C were chosen that were felt to give the best compromise between precision and recall per the criterion stated above for evaluating tests. Details of test results for separate values of C are not included in this work - instead results using the “best compromise” value are discussed.

3) Reduction in feature set size

The distance variable used in the creation of features that seek shifts in language of judgement causes the production of a large number of RST feature types. There was a concern that this might impede learning given a classifier's lack of knowledge of the relationship between these features types. Accordingly the feature creation algorithm was modified to create a single feature where multiple features had previously been created for different values of the distance variable. Values associated with the merged features were added together and assigned to the single combined feature.

This merging reduced the total number of feature types to 544. The number of individual (non-judgemental) features was reduced to 52885 (individual feature values were then scaled between 0 and 1 as was done for non-merged features).

These merged features were tested with LIBSVM's RBK and linear kernels as well as LIBLINEAR. Corresponding tests were *LIBSVM_MERGED_RBK*, *LIBSVM_MERGED_LINEAR* and *LIBLINEAR_MERGED.*

Figure C.1 and Table C.1 now summarise test results.

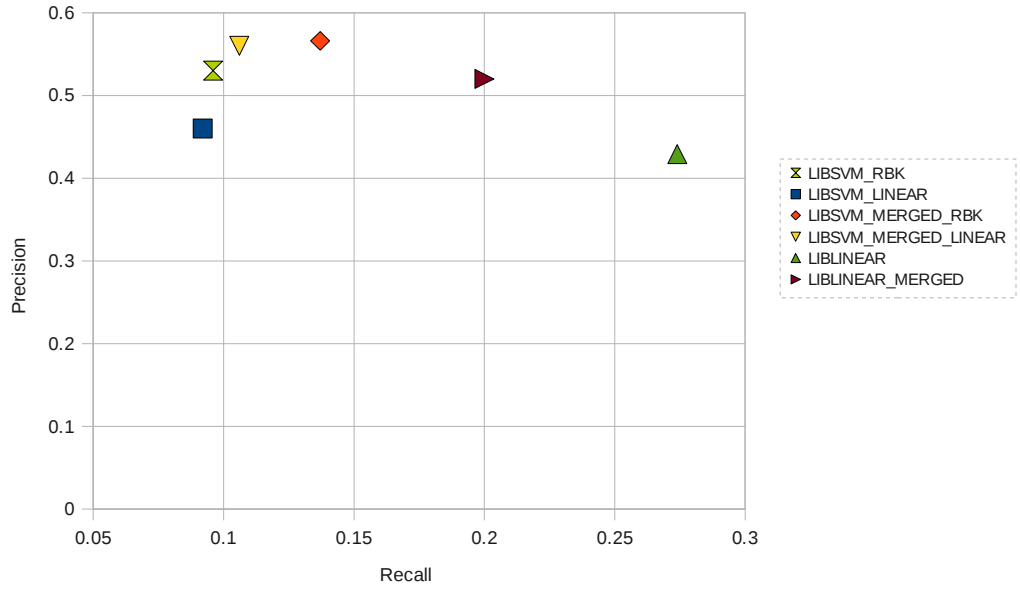

Figure C.1: Relative performance of experiments in acquiring language of judgement

Method	Precision	Recall
LIBSVM_RBK	0.530	0.096
LIBSVM_LINEAR	0.460	0.092
LIBSVM_MERGED_RBK	0.566	0.137
LIBSVM_MERGED_LINEAR	0.560	0.106
LIBLINEAR	0.429	0.274
LIBLINEAR_MERGED	0.520	0.200

Table C.1: Relative performance of experiments in acquiring language of judgement

For LIBSVM, the RBK kernel outperformed the linear kernel for both recall and precision. Datasets containing merged features had higher precision than their non-merged counterparts. Merged features had slightly higher recall than non-merged features when LIBSVM was used but the reverse was true for LIBLINEAR. Tests with LIBLINEAR scored higher values of

recall than tests performed with LIBSVM. Overall, the use of LIBLINEAR with merged features (*LIBLINEAR_MERGED*) was felt to give the best compromise between precision and recall. Accordingly LIBLINEAR was chosen as a classifier and the merged features were used.