Personal sense in subjective language research in the blogosphere

Polina Panicheva

Master of Science (research) in Information Technology

Institute of Technology Tallaght
Dublin, Ireland

2011
Personal sense in subjective language research in the blogosphere

Polina Panicheva

being a thesis presented for the award of Master of Science (research) degree in Information Technology

Supervisors

Dr John Cardiff
Department of Computing
Institute of Technology Tallaght, Dublin, Ireland

Prof Paolo Rosso
Universidad Politecnica de Valencia, Spain

Submitted to the Higher Education and Training Awards Council (HETAC)

May 2011
Abstract

Blogs are a very important part of the digital world, indeed they can be viewed as a digital representation of the whole world. People share pictures and videos, describe their daily life, ask questions and, of course, give opinions. The blogosphere presents a unique opportunity to obtain huge statistics about what people like, feel, need – about their ‘private states’. The vast and ever-growing volumes of ‘bloggers’ and thus, information, demand an automated way of analyzing blog texts.

This gives rise to a new research direction combining computing, linguistics and psychology: sentiment analysis: the computational treatment of (in alphabetical order) opinion, sentiment, and subjectivity in text. Objective characteristics of the writers based on their texts can be analyzed: their age, gender, social affiliation, character; subjective characteristics as moods, negative or positive opinions – polarity, emotions towards an object – can also be investigated.

In the thesis we argue that individual component of word meaning is an essential phenomenon in subjectivity, and elaborate the notion of Personal Sense, described in [Leontev, 1978], to apply it to subjective language investigations. The Personal Sense technique is developed to represent objective characteristics and private states of blog authors. We describe a number of experiments where we demonstrate how more can be said about the author and their private states by harnessing individual Personal Sense; and present a new Personal Sense-based subjectivity, polarity and emotion annotation framework.
Declaration

I hereby certify that the material, which I now submit for assessment on the programmes of study leading to the award of a Master of Science (research), is entirely my own work and has not been taken from the work of others except to the extent that such work has been cited and acknowledged within the text of my own work. No portion of the work contained in this thesis has been submitted in support of an application for another degree or qualification to this or any other institution.

_________________________________________    ________________
Signature of candidate     Date

I hereby certify that to the best of my knowledge all the unreferenced work described in this thesis and submitted for the award of a Master of Science (research) is entirely the work of Polina Panicheva. No portion of the work contained in this thesis has been submitted in support of an application for another degree or qualification to this or any other institution.

_________________________________________    ________________
Signature of supervisor     Date
Acknowledgements

I would like to thank my supervisor, Dr. John Cardiff, for all the guidance and advice I received from him. I am thankful to my co-supervisor, Dr. Paolo Rosso, for his valuable advice.

I would like to acknowledge the support received from Dr Barry Feeney and all the Computing lecturers and technicians.

I would like to thank my colleagues at the ITT Dublin and UPV Valencia, especially Maria Mitina and Fernando Perez Tellez, for their patient help and collaboration. Thanks are also due to Dr Joe McDonagh for his advice and suggestions.

This research work was carried out with the support of a grant from the Institute of Technology, Tallaght, Dublin. Funding for attendance at the LREC Conference was provided by the project MICINN TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

Thank you.
# Table of Contents

1. Chapter 1: Introduction ........................................................................................................... 1

2. Chapter 2: Background ........................................................................................................... 5
   2.1. Personal Sense .................................................................................................................. 5
   2.2. Opinion Mining, Polarity Classification ............................................................................. 7
   2.3. Authorship Attribution ..................................................................................................... 9
   2.4. Authors’ Background Characteristics Identification ....................................................... 11
   2.5. Subjective Language Identification .................................................................................. 14
   2.6. The Personal Sense Approach Applied to Opinion Mining in the Blogosphere ........ 16

3. Chapter 3: Opinion Mining Using a Personal Sense Approach ........................................... 19
   3.1. The Co-occurrence Distribution Approach ....................................................................... 19
   3.2. The Movie Reviews Dataset ............................................................................................ 21
       3.2.1. The Dataset Preprocessing ......................................................................................... 21
   3.3. Experiment: Classification Of the Movie Reviews’ Polarity Using Personal Sense ....... 22
   3.4. Results of the Movie Review Polarity Classification ....................................................... 24
   3.5. Conclusions ..................................................................................................................... 26

4. Chapter 4: Opinion Mining Using Authorship Information and Attribution ....................... 29
   4.1. Motivation: Authorship Attribution ................................................................................. 30
   4.2. The Movie Review Dataset by 10 Authors ........................................................................ 31
   4.3. Experiment 1: Polarity Classification Using Actual Authorship Information ............... 31
       4.3.1. Experiment 1 Results and Analysis ............................................................................. 31
   4.4. Experiment 2: Preliminary Authorship Attribution ......................................................... 32
       4.4.1. Experiment 2 Results and Analysis ............................................................................. 33
   4.5. Experiment 3: Polarity Classification Based on the Modified Dataset Using
       Authorship Attribution ......................................................................................................... 34
       4.5.1. Experiment 3 Results and Analysis ............................................................................. 35
   4.6. Conclusions ..................................................................................................................... 36
Chapter 1: Introduction

Blogs are a very important part of the digital world. Abbreviated from “web logs”, they present web pages which contain personal stories or diaries, texts on a particular topic, sometimes videos, pictures, links to other web pages, and are updated regularly with new information. Starting as personal diaries in 1994, they grew in popularity very quickly. In 2007 Technorati [Technorati, 2007], an international blog search engine, reported more than 110 million blogs.

Blogs can be viewed as a digital representation of the whole world. People share pictures and videos, describe their daily life and ask questions and, of course, give opinions. The blogosphere presents a unique opportunity to obtain a vast quantity of statistics about what people like, feel, need – about their ‘private states’ – without asking them and waiting for their response, but simply by reading their blogs. As a result, interested parties such as advertisement agencies and politicians can find out more than ever about their audience. Of course, the vast and ever-growing volumes of ‘bloggers’ and thus, information, demand an automated way of analyzing blog texts.

This gives rise to a new research direction combining computing, linguistics and psychology: sentiment analysis, or opinion mining. These terms are used differently by various authors, but we follow [Pang et al, 2008] in considering them as interchangeable broad terms for the research direction.
Sentiment analysis, opinion mining or subjectivity analysis is the computational treatment of (in alphabetical order) opinion, sentiment, and subjectivity in text. Different research groups define the area theoretically in different ways, and thus take different practical directions. In [Wiebe, 1994] ‘subjectivity’ is defined as private states, ie. “states … not open to objective observation or verification” [Pang et al, 2008]. This includes opinion, emotions, moods etc. ‘Opinion mining’ is defined in [Dave et al, 2003] as extraction and summarization of product features described in the world wide web, and formulation of an opinion about the product based on these features. ‘Sentiment analysis’ is not defined in a single work, and is considered as a synonym for the broad definition of the former two, the computational treatment of sentiment, opinions, and private states in text.

On one hand, the background for our work is provided by the analysis and automatic classification of objective characteristics of writers based on their texts: their age, gender, social affiliation, character. On the other hand, subjective characteristics such as moods, negative or positive opinions – polarity, emotions towards an object – are also investigated. Thus, based on a number of blog texts by a certain writer and no additional information, their age, gender, personal characteristics, moods when writing certain blog posts and emotions towards the subjects of their discourse can be automatically inferred within a certain accuracy.

Leontev [Leontev, 1978] stated that consciousness is subjective, and defined two types of word-meaning: significance, being the meaning shared by the speakers of a language and representing a part of the objective reality, and Personal Sense, representing subjective characteristics in consciousness, in terms of unique experience of a person. Thus, Personal Sense, serving as a building block for the subjective consciousness, can be harnessed from the writings of bloggers, in order to more accurately deduce information about their opinions, private states and sentiments.

In this thesis, we set of to investigate how Personal Sense can be harnessed for the purpose of blog analysis. Our research is centered around four categories of experiments. We start by exploring Personal Sense reflecting positive or negative polarity in movie reviews. We conclude that the Personal Sense can be successfully applied to polarity classification of products. We then proceed to obtain knowledge of individual writing styles – idiolects – and use this to perform polarity
classification more accurately, thus testing the hypothesis that Personal Sense affects idiolects to a considerable extent and should be exploited in order to perform opinion mining more accurately.

Personal Sense is also used to reflect subjective concept structures specific to a certain professional background. The concept structures are represented as Personal Sense thesauri, and we confirm these to be a useful indicator in personal background classification.

Finally, we develop a new Personal Sense annotation framework, for annotating and classifying subjectivity, polarity and emotion. The Personal Sense framework yields a high performance in a fine-grained sub-sentence subjectivity classification, and provides a promising background in polarity and emotion classification.

The rest of the thesis is organized as follows. In the second Chapter, we present the background for our research, including its theoretical foundations – the discussion of Personal Sense, and existing practical approaches to sentiment analysis, especially the ones related closely to our research direction. In Chapters 3-6 we describe our experiments in sentiment analysis using the concept of Personal Sense. In Chapter 3 we present a co-occurrence based Personal Sense approach to a polarity classification task in opinion mining. In Chapter 4 we introduce an approach to opinion mining that is personalized and thus draws important advantages from authorship attribution. In Chapter 5 a framework for identifying authors' professional background by constructing and classifying their Personal Sense thesauri is presented. In Chapter 6 we describe a lexico-syntactic approach to classifying subjectivity using Personal Sense in news headlines, a Personal Sense framework for subjectivity, polarity and emotion annotations in text is introduced, and results of a subjectivity classification experiment in terms of the suggested framework are presented. Finally in Chapter 7 the conclusions of our experiments are given, and directions for future work are outlined.
In this chapter we describe the theoretical background of our approach. The notion of Personal Sense as a former of the human consciousness sense in studies by Leontev, its close relation to subjectivity and the importance of introducing the concept in the subjectivity analysis field, is discussed in section 2.1. Furthermore, we give an overview of the work undertaken in subjectivity analysis that serves as a background for our studies. In section 2.2 we summarize the work done in opinion mining, with an emphasis on polarity classification in particular. In section 2.3 we give a short overview of authorship attribution. We proceed to the state of the art in identification of the author's background in section 2.4. Section 2.5 contains an overview of subjective language identification. Finally, in section 2.6 we summarize the points made in this chapter and outline an approach we take in our experiments.

2.1 Personal Sense

Our consciousness is subjective. It is subjective in a sense that the objective world is filtered to be perceived in terms of our physical and spiritual needs. In [Leontev, 1978] word-meanings are called “the most important 'formers' of human consciousness”. The dual nature of consciousness is defined by Leontev, as “a picture of the world, opening up before the subject, in which he himself, his actions, and his conditions are included” ([Leontev, 1978]). The author underlines the fact that consciousness is potentially unlimited to reflect objective reality, but is actually determined and thus limited by personal needs, goals and activities.
As the most important former of consciousness, word-meaning is also considered in a dual manner, combining the objective (shared between the speakers of the same language) representation of reality, and the subjective which serves as a building block and an object of individual consciousness. Thus, Leontev suggests a distinction between significance and sense as two types of meaning, exemplifying the distinction with an exam mark: the significance of an exam mark is shared, as everyone who has ever studied knows the meaning of the "exam mark" and its consequences. On the other hand, in individual consciousness an exam mark acquires a certain sense in terms of actual goals of a person, such as advancing their career, impressing those around them for a student obtaining the mark; or being a successful teacher for an examiner; or a decision on how many students stay for a repeat year and how many get a scholarship for a college official, etc. Generalizing this difference, meanings of words in language have a two-fold nature in this respect: a shared and abstract one, and a personal but more actual one. Perception and reflection of the objective reality in individual consciousness is always connected with achievement of personal concrete goals and performing actions, to satisfy their needs, regardless of whether the motives are perceived consciously by the individual or not. The needs and motives make a constant contribution to the filtering of reality in consciousness by evaluating the significance of objects for the individual, thus ascribing Personal Sense to the objects and objective circumstances, in addition to their objective meaning.

Thus, we understand Personal Sense as a component of word-meaning different for each individual, reflecting an object in word-meaning in terms of unique experience of a person. Personal Sense, as word-meaning, is not manifested explicitly in text or speech. Word-meaning in text is analyzed with latent techniques, for example, Latent Semantic Analysis [Pino et al, 2009], with successful applications to various linguistic tasks such as Word Sense Disambiguation [Mitrofanova et al, 2008]. Personal Sense as a characteristic of individual language use should be studied in texts by different individuals separately, taking into account the personality of the authors or their objective characteristics.

An idiolect is ‘a language that can be characterized exhaustively in terms of … properties of some single person at a time’ [Barber, 2009]. An idiolect represents a collection of personal characteristics, i.e., age, gender, social class, occupation, as
well as personal traits and private states. Thus, idiolect can be seen as a combination of the so-called sociolect, genderlect, slang, jargon, etc. Thus, in the same way as implicit word-meaning is studied using different techniques in a language, Personal Sense can be approached in a personal language, i.e. idiolect. This of course does not imply that Personal Sense can only be investigated in texts by a single person; it is rather useful to be aware of the personal characteristics, distinguish texts and authors and compare them in terms of these characteristics. We cannot study idiolect, and Personal Sense in particular, without referring to language and common word-meaning, however the study of idiolect should be the study of language with a more fine-grained analysis of the data.

We find it very important to use the concept of Personal Sense in the linguistic study of subjectivity, opinion and sentiment, as it was defined just as the element that implies subjectivity in word-meaning and serves as a bridge between the objective world and the subjective consciousness. It has become especially important in recent years, as a lot of subjective information has become available on the world wide web: product reviews, blog posts, plagiarized passages, news and opinion texts. This amount of data has never been available for processing in a digital format, which is why we find it important to explore the possibilities of subjective language analysis. We use the Personal Sense construction from a theoretical research work that hitherto could not enjoy such a vast amount of data for analysis. Now the blogosphere grants us the possibility of applying the theory to practice and learning more about the subjective side of personal writings in the web.

2.2 Opinion Mining, Polarity Classification

Research on analysis of blogpost writings is an area attracting an increasing amount of attention. One of the reasons for this is that the blogosphere provides a vast and ever-increasing amount of data characterized by subjectivity of the language used by the authors. Subjective language contains information about private states. One of the popular research domains is the analysis of documents containing the subjective opinion of the author. Much work is also dedicated to the subject of eliciting emotions ([Riloff et al, 2003]), and moods ([Mishne et al, 2006]) of the author.

A well-known task in subjectivity analysis is the polarity classification of documents: texts that contain, for example, product reviews can be automatically divided into two groups: positive and negative reviews. Sometimes a neutral class is
also introduced. Thus, the goal of polarity classification of documents is to ascribe a certain (positive, negative or optionally neutral) class to a document. Interest in this research direction has a long history [Carbonell, 1979], but after 2001, when the period of a fast growth of personal text volumes in the web began, the area has gained high popularity, and now numerous approaches are successfully applied. A broad description of the work done in the area of subjectivity analysis and polarity classification can be found in [Pang et al, 2008].

Various linguistic features are used in opinion mining and polarity classification. The features are traditionally divided into syntactic and lexical. As we are concerned with word-meaning and Personal Sense and how it is represented in text, our work lies in the area of the lexical approach, and as a result of our experiments we provide an efficient lexico-syntactic framework for subjectivity classification using Personal Sense. In this section we will give an overview of the lexical approaches to opinion mining, and mention some syntactic works that are particularly relevant to our approach.

Different lexical features were compared in [Pang et al, 2002]. Word unigrams in various combinations with syntactic information and higher-order n-grams are reported to yield the best results: 82% accuracy in a two-fold classification task. Word-meanings are used usually in their polar or emotional aspects, as for example in [Snyder et al, 2007]. Adjectives, along with other features, such as parts of speech, syntax constructions, the use of negation are the main classes of features used and compared in polarity classification, see [Pang et al, 2008] for numerous examples. In [Riloff et al, 2006] a hierarchy of lexical features is presented, the information gain of different features is discussed, and the hierarchy is employed for the selection of the best features for opinion analysis. In our approach, the features represent information about the position of words in relation to other words, whereas in other works [Pang et al, 2002] information about the position of the words in terms of the whole text is employed, for example if the word is used closer to the middle or to the end of the document.

An approach to opinion classification utilized very successfully is based on product features and their characteristics discussed in the reviews, see [Balahur Dobrescu et al, 2009]. The basic principle of the algorithm is that there are certain features important for a product to be successful, for example in the case of a camera they
would be resolution, size and design – a camera with high resolution, tiny size and nice design is likely to be rated high or positive. However, in our work we are exploiting the task which involves a greater degree of subjectivity in the sense that for other objects, like movies or news, there are no clearly defined features which can have ‘positive’ or ‘negative’ values: for example, a ‘complicated plot’ can be considered an advantage by one movie critic and a shortcoming by another.

Semantic relations like synonymy, hyponymy, causation are well-established in language. Each has certain ways of representation as syntactic constructions in text. A tool called LEILA (Learning to Extract Information by Linguistic Analysis) described in [Suchanek et al, 2006] “learns” such semantic relations from examples, finds the corresponding syntactic constructions and additional pairs of words for these relations. It has been shown to look successfully for such semantic relations as synonymy and date of birth. The LEILA system works as follows. A user defines a relation between words with examples and counterexamples. A syntactic linkage is defined as a syntactic sub-tree, connecting a number of words. The algorithm looks for linkages between the pairs of words. It replaces the words in the linkages by placeholders and produces a syntactic pattern. Then it runs through the text again, finding all the linkages that fit the counterexample pattern. Thus, it gets the negative pattern. Statistical learning is then used to learn the positive patterns, so the pattern classifier is obtained. The program then goes all through the text again, and for each linkage in the text all the possible patterns with particular words are generated. If the pattern for a particular linkage is positive, then the two words connected by that linkage become an output pair of words.

2.3 Authorship Attribution
The main idea behind automatic authorship attribution is that some formal stylometric features can distinguish between texts by different authors. The foundation for the contemporary stylometric approach (making use of the features that in combination describe an author’s writing style) to authorship attribution was laid in [Mosteller et al, 1964], where they investigated the task of attributing ‘The Federalist Papers’. ‘The Federalist Papers’ are 85 articles published in 1787-1788 containing ratification of the Constitution of the United States of America. The articles had been written as a combined work by 3 different authors and have been attributed controversially, becoming thus one of the notorious disputes in the history
of America. Initially such features as sentence length, word length, word frequencies, character frequencies, vocabulary richness and their functions were proposed as reflecting the author’s original style.

The research area has become very popular, and the number of features used nowadays has grown rapidly, along with the vast growth of the number and volume of texts available electronically.

The blogosphere provides vast numbers of texts by different authors. For authorship attribution text by one single author can be used in different ways. First of all, in a profile-based approach, all the known texts by one author are combined into a single document, which it is used to build the author’s profile and extract author-specific features ([Keselj et al, 2003]). Then a text with unknown authorship can be compared to every profile text to find the closest match, using various distance measures. Alternatively, an instance-based approach (e.g., [Sanderson et al, 2006], [Koppel et al, 2006]) considers every text as a classification instance, and draws author-specific features from each of their texts separately.

The authorship attribution problem relates closely to the plagiarism detection problem: identification of unacknowledged reuse by one author of parts of text written by another person. The plagiarism detection methods can be basically classified as intrinsic or external plagiarism detection. External plagiarism detection is particularly similar to authorship attribution, as a reference corpus by different authors is at hand, and the task is to detect plagiarized pieces of text by comparing them to the documents from the reference corpus. Intrinsic plagiarism detection is a much more complicated task, as there is no reference corpus, and the plagiarized passages have to be detected based on differences and inconsistencies between the pieces of the same text. In order to detect the extreme differences, i.e. those that are much bigger than the mean deviation of the overall text writing style, much more accurate and clear features are necessary.

The most widely used authorship attribution features fall under 4 categories: lexical, syntactic, semantic and character. The lexical features are represented by the frequency of word n-grams and single words ([Sanderson et al, 2006], [Keselj et al, 2003]), vocabulary richness ([de Vel et al, 2001]), word and sentence-length ([Yule, 1938]). The syntactic features include parts of speech ([Baayen et al, 1996]),
frequency of structural chunks and rewriting rules ([Gamon, 2004]). The semantic features are semantic dependencies, synonyms or other semantic relations.

The character features require special attention, as in recent years they have been applied successfully to the task. However they do not have an obvious intuitive explanation, as the other mentioned groups of features do. Moreover, the character n-gram features perform successfully in the intrinsic plagiarism detection task [Stamatatos, 2009(2)], and thus are justified as very precise features for building up an author’s profile and performing authorship attribution.

The authors of [Stamatatos, 2009(2)] used character n-grams to perform intrinsic plagiarism detection in the corpus of the 1st International Competition on Plagiarism Detection. They defined a sliding window of 1000 symbols, and compared the text in the window (moving it over the text length) to the features of the whole text, thus getting a function representing the change of the style within the document. The extreme values of the function were supposed to indicate plagiarized passages, as they represented a writing style of a passage that is extremely different to that of the whole document. They reported the result of 0.29-0.31 F-score, with the recall value being 45-46%, and precision 22-23% for the development- and competition-parts of the corpus. We consider it a useful path to follow in plagiarism detection and authorship attribution, and apply a modified algorithm based on the character n-gram features discussed in this work.

2.4 Authors’ Background Characteristics Identification

A task that has grown significantly in popularity recently is the inference of background information about the author from their linguistic behavior ([Pennebaker et al, 2001], [Argamon et al, 2005], [Nowson et al, 2005]). The background information can be divided into personality traits and objective background, which includes age, gender, and native language of the author, manifesting in the sociolect of the author. The personality traits are a more complex phenomenon to be classified, as it requires an additional stage, where the meaning of a personality trait is defined according to a theory of emotions, and respectively a psychological test is applied to obtain a gold standard for the classification. A sociolect is a language shared by a relatively small group of people sharing the same objective feature: age, gender, geographical, educational or career background. In the sociolect identification the gold standard is, on the one hand, mostly provided with the text,
when we know the age and gender of the author. Even in the cases in which these are unknown, a questionnaire can be applied which requires no additional elaboration and raises no questions about its theoretical congruity.

Various statistic word-count measures have been applied to identify both personality and background characteristics of the author. All of them are based on the assumption that presence or high frequency of certain words are indicative of a respective trait or class, whereas their absence or low frequency indicate a different, perhaps opposite, class or trait value.

The authors of [Pennebaker et al, 2001] introduce a software tool, Linguistic Inquiry and Word Count (LIWC), designed for statistical analysis of word usage. The tool offers a huge variety of options, for instance one can choose from counting the average number of words in a sentence, to the frequency of the words which indicate positive emotion, optimism or activity. In [Pennebaker et al, 2003] they give a short overview of the state of the art in finding correlations between language use and personality from the psychology point of view.

Recently the “Big Five” ([Goldberg, 1990]) personality traits have become the most popular among linguistics researchers, probably because language vocabulary itself played an important role in their formulation ([John et al, 1999]). Most of the work in personality trait identification is made based on different combinations of the five features. For example, in [Argamon et al, 2005] a two-scale personality trait text classification is performed: Neuroticism and Extraversion are identified. The corpus consisted of stream-of-consciousness and deep self-analysis essays written by students. A text is represented by a set of around 600 function words, and semantic lexical attributes related to cohesion, assessment, or modality, and appraisal, or judgments about quality. The model was trained with the Sequential Minimal Optimization [Platt, 1998] algorithm. For classification on the scale of Neuroticism, the accuracies for both types of texts are almost the same, and quite modest: 50% for function word counts and 58% for the best feature – appraisal, which is not surprising. As expected, high Neuroticism is associated with negative appraisal. For Extraversion the scope of the accuracy is the same (50-58%), the best values are for

---

1 The modern understanding of the Five Factor Model names the following personality features: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.
all of the features together (with stream-of-consciousness texts) and for function words counts (for self-analysis texts).

In [Oberlander et al, 2006(1)] the authors use a corpus of students’ personal e-mails to identify the connection between implicitness in language behavior and Extraversion and Neuroticism. In [Oberlander et al, 2006(2)] they investigate all five personality traits, using n-grams as features, and proceed to test the resulting model on a larger corpus in [Oberlander et al, 2007].

Among the characteristics of the author forming a sociolect, gender is broadly investigated. In [Nowson et al, 2005] the authors introduce the notion of formality versus contextuality, contextuality meaning high degree of deixis and such parts of speech as verbs, pronouns, adverbs and interjections, and formality - nouns, adjectives, prepositions and articles. In non-academic speech, men are found to prefer more formal style of communication, while women use a more contextual style.

The authors of [Argamon et al, 2003] also study text classification by the gender of the author. The training features included function words, and part-of-speech n-grams. The authors conclude that women tend to communicate in a more subjective manner, while men in a more objective one.

The study described in [Schler et al, 2006] is concerned with age and gender classification of blog texts. Four types of text stylistic features are chosen as follows: selected parts of speech, function words, blog neologisms and hyperlinks. According to their age, the bloggers were divided into three categories, and with the whole set of features giving the best result, the classification performance reached 75% accuracy; whereas the entire set of features for the gender classification performed with 80% accuracy.

Another important category in authors’ sociolect identification from their texts is native language. Authors with different native language background may write in the same language using it differently: they prefer certain syntactic constructions or make different mistakes. In [Koppel et al, 2005] some stylistic features of text are used to identify the author’s native language. These are function words, letter n-grams, various types of mistakes and idiosyncrasies. They analyze texts in English
by native speakers of 3 Slavic languages, Spanish and French. When all feature types are used together the accuracy obtained is as considerable as 80.2%.

In our work we investigate the professional sociolect of blog authors based on the identification of Personal Sense interrelations in their texts. Our research also applies to the scope of perspective determination: identification of the author’s social, ideological, professional background based on their texts. It is a new and vast-growing research area.

In [Choudhury et al, 2008] the authors present an approach to perspective determination based on concept interrelation ontology. They manually construct an ontology of the movie production field. The semantic relatedness measure for 25 pairs of words denoting motion picture industry concepts is presented. The next step suggested is gauging the author’s perspective within the motion picture industry field.

The authors of [Yoshida et al, 2003] describe a method for constructing a personalized thesaurus from bookmarked web pages and documents. They construct a personalized thesaurus for each user as a context distributional profile of a word and propose a scale for measuring the difference between the thesauri.

2.5 Subjective Language Identification

Subjective language is language used to express private states ([Wiebe et al, 2004]). Subjective language identification is an area in Opinion Mining aimed at discovering pieces of text containing subjectivity, i.e. information about private states, without trying to understand the meaning of the information. Consider the following example from a movie review:

“The movie really does have a very interesting plot line and is a true story about a person that I personally knew nothing about.” (s1)

EB from the Junkies and I went and saw the film and both agreed that even know we didn't know who the guy was, it was really interesting to learn about him, considering how he important he was to the drug world.” (s2)

The film takes place in the late 60s, early 70s and tells the story about an American gangster named Frank Lucas. Lucas (played by Washington) was…” (s3)
It is clear that in the first two sentences the author lets us know about their attitude towards the film: they find it interesting, and in the following piece they proceed to describing the plot of the movie. The goal of a subjective language identification system here would be to identify the sentences s1 and s2 as subjective, i.e. containing an opinion and/or an emotion, without going further to classify a positive or negative opinion polarity or a certain emotion of interest, joy etc.

Here the goal is to ascribe a subjective or objective class to a sentence, thus the analysis would aim at sentence-level subjectivity identification. On the other hand, much research work is dedicated to document-level subjectivity identification [Spertus, 1997]. In our work the analysis will be focused on a fine-grained subsentence-level, with the main classification element being a pair of words.

Subjectivity is identified with subjective clues. These are lexical items (words or collocations) which contain subjectivity and attach it to the analyzed sentence or document. First of all, a vocabulary of subjectivity clues is constructed (see, for example, [General Inquirer, 2000]). The vocabulary clues are in turn identified in text and used to predict subjective pieces [Wiebe et al, 1999]. Sometimes the subjective clues can also be used objectively [Wiebe et al, 2004], in which case a disambiguation algorithm is applied. Finally, a sentence or document is identified as subjective or objective using the selected lexical and probably some additional features and a classification algorithm. For instance, the authors of [Yu et al, 2003] and [Wiebe et al, 2004] apply the described subjectivity identification scheme and report results of 96.5% F-measure and 94% accuracy respectively, in a document-level classification.

In our research on subjectivity identification we identify the subjectivity clues using existing affective vocabularies. The syntactic constructions containing the clues play a unique role in our work: they are used among other features for the clues disambiguation, and serve as patterns for regular subjectivity constructions extraction. Thus, especially relevant to our work is the syntactic approach to opinion mining based on extraction patterns [Riloff et al, 2003]. A news text dataset consisting of 34,000 sentences is investigated in the work. The authors apply an algorithm to learn syntactic patterns associated with subjectivity. The resulting patterns are bootstrapped in order to classify sentences as subjective or objective.
In our experiments we are using a similar syntactic pattern extraction approach, but we apply it in a new object-oriented framework to classify word pairs.

2.6 The Personal Sense Approach Applied to Opinion Mining in the Blogosphere

We have described various directions in subjectivity analysis, which have recently become, because of the vast growth and availability of personal texts in the blogosphere, a very popular and fruitful area of research. In polarity classification, authorship attribution, background characteristics identification of authors, and in the basic subjectivity identification, various features have been proposed and high results achieved. However, it is never mentioned that a personal component of meaning – Personal Sense – is a very important feature of any subjective text, moreover, it forms unique idiolect features and reflects personal preferences in text; although in theory its role in subjective language research is obvious, as it is defined as a former of subjective consciousness.

In polarity classification the Personal Sense of an appraised object may reflect the appraisal as a part of the word-meaning. Furthermore, different individuals choose different ways of expressing appraisal in reviews and blog texts, and a characteristic that is good for one person can be bad for another one. Thus it forms a unique writing style for each author, which is important to take into account when we analyze text polarity on the one hand, and on the other hand can be used to infer background experience automatically. Finally, the notion of Personal Sense provides a possibility to annotate and identify emotions towards specific objects in a fine-grained manner, learning regular patterns of emotional expressions.

We find it important to explore the opportunity provided by the vast amount of personal texts in the web and the notion of Personal Sense in subjectivity analysis. Thus we set out to harness the Personal Sense of bloggers, in order to analyze their texts both individually and collectively and be able to say something about bloggers.
Chapter 3: Opinion Mining Using a Personal Sense Approach

Our approach to opinion mining is based on the detection of Personal Sense. In this chapter we apply the co-occurrence distributional technique (used to model word-meaning in many applications) to the individual component of word-meaning: Personal Sense. Our hypothesis is that the Personal Sense of words, detected with this technique, represents the author’s opinion about the object or event denoted by the word, i.e. polarity: a positive or negative appraisal of the object. We analyze polarity of movie reviews with the suggested technique. The accuracy of our results slightly outperforms the baseline. We discuss the reasons for the modest performance of our method.

3.1 The Co-occurrence Distribution Approach

The co-occurrence distribution method has been applied to various tasks in word-meaning studies. Co-occurrence of a word is the words or other elements that occur together with the current word. The context (or “company”) of a word is represented by its distributional profile [Mohammad et al, 2006]. The technique is based on the distributional hypothesis: an assumption that a word-meaning is represented in the word usage [Wittgenstein, 1973], [Rubenstein et al, 1965]. “You shall know a word by the company it keeps” ([Firth, 1968]). A widely used example ([Mohammad et al, 2006]) is that of the word 'bank'. Encountering this word in a text, its sense is identified using the context: if it is surrounded by the words 'river', 'silt', 'water', we realize that most probably it is a river bank that is meant. On the other hand, if the context includes 'money', 'account', 'robbery', we identify that 'bank' is currently used
in the sense of 'financial institution'. Thus a word meaning can be represented by a set of its contexts and compared to the meanings of other words. The co-occurrence distribution technique is applied, for example, in the LSA technique ([Rohde et al, 2009]), or in Word Sense Disambiguation ([Mitrofanova et al, 2008]).

The co-occurrence distribution technique could be described as follows. First of all, each word in question (the ‘target’ word) is represented by a vector, combining all of the words contexts occurring in the corpus. Every dimension in the multi-dimensional vector space is represented by a word occurring in a target word’s context. In our work, the ‘context’ means a number of words occurring immediately before and after the target word. The occurrence count of the current dimension word in the context becomes the value of the vector on the current dimension. Thus for every target word we get a vector in a multi-dimensional vector space, and can measure proximity between the vectors that would correspond to our intuition that the more the word contexts have in common, the closer the vector lie to each other. Following the distributional hypothesis, the higher proximity the vectors have to each other, the closer the word meanings are, and the more likely the target words are to be synonyms [Rubenstein et al, 1965]. For example, we consider the following sentence

“I went to see the movie with a friend” (ex. 3.1)

and build a co-occurrence distributional profile of the word 'movie' for different context window length: 1, 2 and the whole sentence (see Table 3.1).

<table>
<thead>
<tr>
<th>Context window</th>
<th>Words representing vector dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>I went to see The movie with a friend</td>
<td>1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0</td>
</tr>
<tr>
<td>I went to see The movie with a friend</td>
<td>1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0</td>
</tr>
<tr>
<td>I went to see The movie with a friend</td>
<td>1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0</td>
</tr>
</tbody>
</table>

Table 3.1. Examples of co-occurrence vectors

Personal Sense is an individual constituent of word meaning. Thus if a word is given that is used by several authors in their texts in the same meaning, and the co-
occurrence, or contexts, of the words are different in texts by different authors, we infer that this represents difference in the personal aspect of meaning, i.e. the Personal Sense.

Thus we have applied the co-occurrence vector space technique to represent the meaning of the words ‘movie’ and ‘film’ and capture the different Personal Sense of these words in positive and negative movie reviews.

3.2 The Movie Reviews Dataset

We have used the movie-review corpus introduced in [Pang et al, 2002]. The corpus contains 1400 short movie-reviews, 700 of them positive and 700 negative. It is obvious in the manual examination of the corpus, that in the case of movie reviews, there are no common features that can be rated bad or good for the movie. On the other hand, sometimes the polarity of the review is opposite to the values of the features: ‘it's the story that needs some major work’ – ‘major work’ would be indicative of a positive appraisal, but the sentence is taken from a negative review, whereas ‘bresson employed nonprofessional actors who recite the dialogue in emotionless, flat voices’ is a piece of a positive review.

3.2.1 The Dataset Preprocessing

We use minimal text preprocessing: we turn all the letters to lowercase, and delete stopwords. We use the obtained text collection for the first experiment. In the next experiment we use some initial keywords, namely ‘movie’, ‘film’, and both of them at a time. For our experiment we required every review file to contain the keyword. Moreover, the experiment results appear to be more reliable, the more times the keywords occur in the review files [Mitrofanova et al, 2008]. For the initial experiments we set the minimum possible occurrence of keywords to 2. That means we only take into account the texts that contain the keyword two or more times. Thus we formed groups of files with positive and negative reviews for the keywords ‘movie’, ‘film’ and for both – ‘movie’ and ‘film’; in the latter case using the files that contained either ‘movie’, or ‘film’ twice, or once the ‘movie’ and once the ‘film’.

We expected the frequency of the keywords to be crucial for our experiments, so for comparison we constructed a corpus from the texts that contained the keywords at least 4 times only. In terms of frequency this is still a very small number, but it is
suitable for our current goals, as we are testing the performance of the co-occurrence method with short review texts, and with the current corpus this threshold allows enough text instances.

In Table 3.2 the numbers of files appropriate for our task containing different polarity reviews for each of the keywords are presented.

<table>
<thead>
<tr>
<th>occurrence &gt;= 2</th>
<th>movie</th>
<th>film</th>
<th>movie_film</th>
</tr>
</thead>
<tbody>
<tr>
<td>pos</td>
<td>557</td>
<td>415</td>
<td>671</td>
</tr>
<tr>
<td>neg</td>
<td>525</td>
<td>439</td>
<td>649</td>
</tr>
<tr>
<td>occurrence &gt;= 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pos</td>
<td>381</td>
<td>198</td>
<td>565</td>
</tr>
<tr>
<td>neg</td>
<td>353</td>
<td>254</td>
<td>568</td>
</tr>
</tbody>
</table>

3.3 Experiment: Classification Of the Movie Reviews’ Polarity Using Personal Sense

We are testing the hypothesis that word-meaning, and especially Personal Sense, is represented between the words, in the way that words are put together to form a text and in the way they occur together with each other. We use a simple co-occurrence matrix to represent word-vectors. Then for the word that we examine we obtain a list of the most similar word-vectors, with the similarity measure being the cosine measure between the two vectors [Mitrofanova et al, 2008]. We assume that the list of the most similar word-vectors characterizes the Personal Sense of the word for the author in the particular text.

Here we do not address the question of which words, occurring in the similar word-vector lists, would reflect the positive or negative component of the Personal Sense. Moreover, we suppose that every critic will use different words to reflect positive or negative appraisal, although for some reviews the co-occurrent word-lists are clearly indicative of the author’s sentiment. Consider, for example, the list of the most similar words to the trigger-word ‘movie’: [series desperate induced climax talking]
fourth abruptly short theater hush bad] drawn from a negative review, or [typical worth reinterpretation unsure essentially inspired attempt copy beautifully lush hollywood angels films early make desire stevens striking live] from a positive one. Our goal is to check the following hypothesis: do the most similar word-vectors reflect the appraisal component, and thus improve the polarity classification results?

Our hypothesis is that the Personal Sense of the word contains an appraisal part. In the case of positive and negative movie reviews, the Personal Sense of the word ‘movie’, for example, would be slightly different in each individual review. We believe that the component of positive or negative appraisal would be contained in the Personal Sense of the word ‘movie’, and that this would be reflected, in some subtle way, in the list of the most similar vectors to the word-vector ‘movie’, ie. reflecting the author’s positive or negative experience in text. A person who likes a movie that they have just seen would use the word denoting the movie in a different way to that of a person who does not like the movie.

We performed classification experiments using 10-fold cross-validation with the movie-review corpus. The task was to classify every individual review as negative or positive, with the gold standard provided. For the experiments we used a machine learning tool, Weka 3.6.1\(^2\) described in [Witten et al, 2005].

We used the list of all the unigrams contained in the review as the baseline feature set. The algorithms that we applied were Naïve Bayes (NB)\(^3\) and Linear Support Vector Machine (SVM)\(^4\). Using the overall corpus and the two sets of options described later in Section 4.2, the initial baseline accuracy results were 77.39% for NB and 81.73% for SVM.

For the next stage of experiments we used the groups of files selected for the words ‘movie’, ‘film’, and both. We used each of these three as ‘trigger-words’. First, for the trigger-word we obtained the lists of co-occurrence vectors with similarity higher than 0.20, one word-list for every review-file. At this stage we only used these co-

\(^2\) http://www.cs.waikato.ac.nz/ml/weka/

\(^3\) An algorithm implemented in a Weka class, described at http://weka.sourceforge.net/doc/weka/classifiers/bayes/NaiveBayes.html

\(^4\) http://www.support-vector-machines.org/SVM_soft.html
occurrence lists. We established empirically that a higher threshold (over 0.25) leaves no vectors similar enough to the triggers, for many files in the corpus. A lower threshold (lower than 0.15), on the other hand, gives exactly the same results as the initial 0.20 for some test experiments.

We conducted two groups of experiments: the 'similarity' and the 'presence' experiments, the difference between them being the way we approach the values in the vector dimensions. In the ‘similarity’ experiments, we accounted for the information on actual similarity measures for the words in the lists. In the ‘presence’ experiments, on the other hand, we assign a formal 1.0 similarity to each word in the list, so it is only the presence or absence of a word in the list that makes a difference.

To obtain word-vectors, we also experimented with different options. First, we used different window-size: 1 or 2 words before and after the trigger-word. The second option was to include the words separated from the keywords by punctuation marks in the contexts, or to leave out such words, i.e., to ignore punctuation (-) or to take it into account (+), respectively.

### 3.4 Results of the Movie Review Polarity Classification

Our results are shown in Table 3.3. The numbers in italics are the best results that we achieved for one constant set of options. Thus, the best options for NB were windowsize 1, using punctuation marks, and only accounting for the presence of the similar word-vectors in the co-occurrence lists. For SVM, on the contrary, the outstanding best options were windowsize 2, no punctuation borders, and similarity instead of just presence. For brevity we call them (set 1) and (set 2) respectively in our following experiments.

We used two types of datasets: the texts containing the trigger-words at least twice, and the texts containing them at least 4 times. Table 3.3 only shows results for the first dataset: the second dataset confirms the results obtained for the first one, i.e. the results are similar, and the options yielding the best results are the same.

In our second experiment we combined the baseline features with the co-occurrence features. We used the same sets of review-files. We compared the results for the combination of features with the results of applying the baseline features only to the same sets of files. Table 3.4 shows the second stage results in comparison. Again,
the results for the first and the second datasets were similar, so we only present the results for the first one here.

Table 3.3. Accuracy results for the co-occurrence features.

<table>
<thead>
<tr>
<th></th>
<th>film</th>
<th>movie</th>
<th>movie film</th>
<th>window</th>
<th>punctuation</th>
<th>expt type</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>55.91</td>
<td>56.24</td>
<td>54.36</td>
<td>2</td>
<td>-</td>
<td>similarity</td>
</tr>
<tr>
<td></td>
<td>51.45</td>
<td>53.46</td>
<td>52.53</td>
<td>2</td>
<td>+</td>
<td>similarity</td>
</tr>
<tr>
<td></td>
<td>51.84</td>
<td>53.24</td>
<td>54.98</td>
<td>2</td>
<td>+</td>
<td>presence</td>
</tr>
<tr>
<td></td>
<td>56.40</td>
<td>57.50</td>
<td>53.86</td>
<td>1</td>
<td>+</td>
<td>presence</td>
</tr>
<tr>
<td></td>
<td>52.51</td>
<td>54.39</td>
<td>50.22</td>
<td>1</td>
<td>+</td>
<td>similarity</td>
</tr>
<tr>
<td>SVM</td>
<td>59.25</td>
<td>59.45</td>
<td>59.97</td>
<td>2</td>
<td>-</td>
<td>similarity</td>
</tr>
<tr>
<td></td>
<td>52.87</td>
<td>57.87</td>
<td>58.51</td>
<td>2</td>
<td>+</td>
<td>similarity</td>
</tr>
<tr>
<td></td>
<td>52.68</td>
<td>58.23</td>
<td>56.39</td>
<td>2</td>
<td>+</td>
<td>presence</td>
</tr>
<tr>
<td></td>
<td>54.27</td>
<td>56.28</td>
<td>53.66</td>
<td>1</td>
<td>+</td>
<td>presence</td>
</tr>
<tr>
<td></td>
<td>56.31</td>
<td>56.23</td>
<td>53.00</td>
<td>1</td>
<td>+</td>
<td>presence</td>
</tr>
</tbody>
</table>

It is obvious that for the same set of files the co-occurrence method makes little any difference in comparison to the baseline. We expect that the two possible reasons for this are: the trigger-words in both cases were too infrequent, so their co-occurent vectors were actually random words and did not really characterize the Personal Sense of the trigger-words, thus the overall review polarity.
### Table 3.4. Accuracy Results for the combinations of features and for the baseline.

<table>
<thead>
<tr>
<th>Experiments</th>
<th>NB</th>
<th>SVM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>options set 1</td>
<td>film</td>
<td>movie</td>
<td>movie_film</td>
</tr>
<tr>
<td>unigrams+</td>
<td>76.99</td>
<td>78.41</td>
<td>78.40</td>
</tr>
<tr>
<td>co-occurrence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unigrams</td>
<td>76.94</td>
<td>78.42</td>
<td>78.39</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>options set 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unigrams+</td>
<td>60.54</td>
<td>63.72</td>
<td>74.05</td>
</tr>
<tr>
<td>co-occurrence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unigrams</td>
<td>76.94</td>
<td>78.42</td>
<td>78.39</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The set of files containing at least 2 keywords itself yielded better performance than the overall corpus, with the overall baseline being 77.39%. This leads us to the intuition that the texts containing frequent explicit names of the subjects are easier to classify in terms of polarity, at least using the unigrams as features.

### 3.5 Conclusions

We have applied the co-occurrence distributional technique to model Personal Sense. Our experiments have confirmed the hypothesis that, to a certain extent, the Personal Sense of words described with the suggested technique, represents opinion polarity and may be successfully applied to automatically analyze polarity in movie review texts.

Our experiments show that first of all, for all the features and datasets used, Linear SVM consistently outperforms NB, and this confirms some of the results in [Pang et al, 2002]. Moreover, NB and SVM with the co-occurrence features show their best performance with totally different text processing options. In other words, the best
performance of NB with the co-occurrence features alone involved punctuation, windowsize 1 and word presence, in all of the options SVM ‘prefers’ the ones that give more information to the features: no punctuation as limits, larger windowsize and real-number word similarity instead of the binary presence.

Secondly, the accuracy for the Personal Sense features alone was in most cases clearly better than the 50% baseline, while it does not significantly improve the baseline in combination with the unigrams. The co-occurrence features obviously contain some quantity of information about the polarity of the reviews, but the same information, and more, is present in the unigram features.

Notwithstanding our expectations, the co-occurrence method in all cases changes the performance very little. The reasons for this we believe to be as follows. Firstly, for each author the Personal Sense of a good movie and a bad movie is different, so the co-occurrence lists for different authors might include different words with no overlap between the authors. This means that the Personal Sense should be first of all investigated for every author individually. Therefore, the next step of the Personal Sense investigation will include polarity classification of the reviews, with texts by different authors being separate datasets.
Individuals write about good things and bad things in their own unique style. In order to analyze polarity of texts automatically, we need to account for this difference. To date, polarity has been analyzed using common features independently of the author. Consider the following samples from movie reviews.

- "It's overly sentimental and at times terribly mushy, not to mention very manipulative. But oh, how enjoyable that manipulation is." (ex. 4.1)

- "Alas, «My Giant» is a film that uses manipulative sentimentality so frequently and with such high intensity that I forgot as I watched it that there are other ways of getting audience members choked up." (ex. 4.2)

Both authors use the apparently 'negative' terms 'sentimentality' and 'manipulative' to characterize the movies – the subjects of their discussion. However, the statement in example 4.1 is a positive review sample, while ex. 4.2 is a negative one. This is confirmed by our intuition reading these short samples alone. The author of example 4.1 justifies the term 'manipulative' in the second sentence of the passage, attaching a positive Personal Sense to this word in their idiolect, and the author of example 4.2 emphasizes its negative effect and Personal Sense towards the end of the sentence. We can conclude that in the idiolect, or individual language style, of the first author the word 'manipulative' is not a crucially negative term, as opposed to the second author. We find this information important in analyzing polarity automatically. One of the basic ways to apply such information is learning and classifying opinion in
reviews by different authors separately. Using examples 4.1 and 4.2 above, separating the subsets for polarity classification would mean that the model would learn 'sentimentality' and 'manipulative' as a negativity clue in the idiolect of author #2 and apply it to classify their texts, but in the case of author #1 the slightly positive Personal Sense of the word in their idiolect would be taken into account, and it would not serve as a clearly negative feature classifying their texts.

In this chapter we suggest a personalized procedure of opinion analysis. Our hypothesis is that Personal Sense influences writing style to such a degree that by analyzing text collections by different authors separately, word-count information in positive and negative opinionated texts can be utilized more accurately to predict opinion polarity. Furthermore, we test a stronger consideration, that if we have no authorship information at hand, application of a sufficiently accurate authorship attribution method may increase the polarity classification accuracy. Thus, we set out to apply polarity classification to different collections of texts. First, we analyze the texts randomly, without taking account of their authorship. Secondly, we divide texts into collections belonging to different authors. Finally, we model a more real-life task: assuming that we have limited information on the text authorship, we apply an authorship attribution algorithm, and test whether, and to what extent, the resulting automatically derived authorship classes can be used to perform the similar polarity classification experiment. The outcome of the experiments confirms the hypothesis, although the applied authorship attribution method imposes certain limitations on dataset processing.

4.1 Motivation: Authorship Attribution

Authorship of documents is not normally taken into account in opinion mining. Given two sets of documents by Author1 (A1) and Author2 (A2), we find some features that A1 and A2 will have in common for both positive and negative appraisal documents. Our consideration is that features that distinguish positive and negative opinion for A1 and A2 may be different. An idiolect, being ‘a language that can be characterized exhaustively in terms of … properties of some single person at a time’ [Barber, 2009], may represent a way to describe personal likes or dislikes. The Personal Sense of a good and a bad movie may be different for individual authors. In other words, A1 and A2 might express and describe their appraisal in an individual way, and their idiolects may be so different, that the overall polarity
classification results might be improved by classifying the documents written by A1 and A2 separately, using the A1 and the A2 set of documents as two different datasets for the experiment. It is not our goal here to investigate the Personal Sense as it is represented in the reviews, but by considering the polarity of the reviews by different authors separately, we implicitly address the Personal Sense of a good and a bad movie for each individual author.

4.2 The Movie Review Dataset by 10 Authors

We have constructed a corpus\(^5\), in a similar format to the one described in [Pang et al, 2002]. It also contains short movie-reviews and is balanced against polarity, but is different in the sense that every author is represented by a considerable number of documents. The corpus consists of 300 short movie reviews, 30 reviews for each of 10 authors. For each of these 30, 15 are positive and 15 are negative reviews. To investigate different corpus volumes and to achieve higher statistically significant results, we have doubled the corpus for each author, and repeated the experiment with 600 documents, 30 positive and 30 negative reviews for each author out of 10.

4.3 Experiment 1: Polarity Classification Using Actual Authorship Information

The aim of the experiment is to observe the hypothesis that in the idiolects of the different authors their appraisal is expressed in an individual way. In order to prove this, we used the unigram features with the Linear Support Vector Machine algorithm for the polarity classification.

As a baseline, we divided the corpus into 10 groups, each having 15 positive and 15 negative (30 positive and 30 negative for the doubled corpus) reviews, in a random way, so that every group consisted of documents by different authors. We performed the 10-fold cross-validation experiment for each of these groups separately. The mean accuracy result for the baseline experiment with the 10 shuffled groups was 56.47% for the smaller corpus, 64% for the doubled corpus.

4.3.1 Experiment 1 Results and Analysis

For the experiment we used the same reviews, but we organized the 10 groups of documents so that each group corresponded to a single author, i.e., it consisted of 15

\(^5\) Available on request at ppolin86@gmail.com.
positive and 15 negative (30 positive and 30 negative respectively) reviews by the same author. The authors were different for every different group. The number of documents and the settings of the classification experiment stayed the same.

The mean accuracy result for the 10 groups was **69.67%**, for the bigger corpus the mean value reached **74.97%**. The t-test (Chapter 5 in [Manning et al, 1999]) showed that for the experiment with 15 reviews written by a single author, the result was better than for the baseline with 15 reviews by random authors, with 75% significance; whereas for 30 reviews by a single author yielded better results than for the baseline, i.e. 30 reviews by different random authors, with 89% significance.

Using the entire corpus as a dataset for the classification, for the 300 reviews the accuracy result was **73.17%**, while for the 600 reviews it was **78.35%**. We summarize the results in Table 4.1.

<table>
<thead>
<tr>
<th></th>
<th>Corpus 1</th>
<th>Corpus 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>300 texts</td>
<td>600 texts</td>
</tr>
<tr>
<td>Overall corpus</td>
<td>73.17%</td>
<td>78.35%</td>
</tr>
<tr>
<td>classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 random groups</td>
<td>56.47%</td>
<td>64.00%</td>
</tr>
<tr>
<td>10 author groups</td>
<td>69.67%</td>
<td>74.97%</td>
</tr>
</tbody>
</table>

The results for the both datasets confirmed our assumption that in different authors’ idiolects the appraisal polarity is expressed in an individual way. Moreover, the mean result for the 60 documents by every author was slightly better than the result for all the 300 documents. This suggests that in terms of the volume of the datasets for polarity classification in the web, it is more useful to double the corpus by the same single author, than to increase it 5 times using texts by different authors.

### 4.4 Experiment 2: Preliminary Authorship Attribution

Knowing the authorship of the reviews, we can use such information and increase the performance of polarity classification. However, this is not a very realistic state of affairs when we use the ever-changing world wide web as a corpus. A very popular way of handling this issue is automatic authorship attribution. In our next
experiment we applied an authorship attribution algorithm to the existing document corpus, investigated if the resulting authorship information increases the performance of polarity classification, and observed the drawbacks and limitations of the approach.

We used the Java Graphical Authorship Attribution Program (JGAAP), described in [Juola et al, 2006], for the supervised authorship attribution task. The tool allows for the choice of the classification features, including lexical, character, phonetic, grammatical features; and the choice of the classifying algorithm: the traditionally used NB and SVM and a number of others. For an authorship classification experiment using JGAAP it is necessary to have at least one training example per author. It starts with learning authorship classes from a trial set of documents by known authors, and proceeds to classifying each document with unknown authorship against the resulting authorship classes.

In our corpus we had collected the 600 documents by 10 authors and a smaller 300 document corpus, both balanced in terms of polarity and authorship. To perform authorship attribution, we used the smaller corpus as a reference group of documents with known authorship: for each author we had a learning set of 15 positive and 15 negative documents. We used the rest, i.e. the second half of the bigger corpus, as a test set.

4.4.1 Experiment 2 Results and Analysis

After testing a number of features: character and word n-grams - the Character Trigrams yielded the best result, confirming our expectation based on [Stamatatos, 2009(2)]. The Cosine distance was experimentally used as the classifier algorithm.

The authorship classification accuracy results for different authors ranged very considerably from 0.3 to 0.93, with the standard deviation of 0.26, and the mean accuracy rate among the 10 authors reaching 0.64. We consider this a successful result, being very close to some of the mean accuracy results reported in [Juola et al, 2006] for the “Ad-hoc Authorship Attribution Competition” (AAAC) ([Juola, 2004]), namely 0.65. On one hand, our reference group contained a large number of documents compared to the AAAC competition tasks, which made the classification task easier. On the other hand, most of the tasks in the competition only included 3
or less known author classes [Juola, 2004], whereas in our case their number was 10, which made the task harder and significantly decreased the task baseline.

The correlation coefficient for polarity classification and authorship attribution results for the 10 authors reached a small but positive number of 0.176, indicating an insignificant trend that the style of the authors that is distinctive in terms of idiolect bears a lot of idiolect features that distinguish it from other authors' idiolects, allowing also for easier polarity classification.

We used the authorship attribution results described in section 3.2 in order to modify our corpus. We applied the results of the classification, so that for each author their document collection contained the documents, whose authorship was considered unknown and was identified automatically by the classifier. Thus, we got 20 collections, a positive and a negative set for each author. Our goal was to proceed with the polarity classification experiment on the new, automatically attributed dataset, to find out if authorship attribution algorithms can contribute to sentiment analysis results, as was the case when authorship was known.

4.5 Experiment 3: Polarity Classification Based on the Modified Dataset Using Authorship Attribution

According to our presuppositions, the application of the authorship attribution algorithm raises real-life issues crucial for the polarity classification task. First of all, the resulting dataset was not balanced in terms of authors. With 30 documents per author at the start, the resulting collections ranged from 8 to 46 texts. Secondly, and more importantly, the collections were not balanced against polarity (see Table 4.2).

<table>
<thead>
<tr>
<th>Author id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>20</td>
<td>15</td>
<td>1</td>
<td>20</td>
<td>15</td>
<td>20</td>
<td>15</td>
<td>6</td>
<td>11</td>
<td>27</td>
</tr>
<tr>
<td>Negative</td>
<td>26</td>
<td>6</td>
<td>7</td>
<td>23</td>
<td>14</td>
<td>23</td>
<td>12</td>
<td>4</td>
<td>14</td>
<td>21</td>
</tr>
</tbody>
</table>

In order to overcome these issues, we supplemented the resulting datasets with documents used as authorship attribution learning examples, so that for each author:

- the volumes of the positive and negative datasets were the same;
- the volume of the set was exactly 30 documents.
For example, for the smallest set, consisting of 1 negative and 7 positive reviews, we added 14 randomly selected negative and 8 positive reviews from the learning set of the same author in the authorship attribution experiment. For the largest set, containing 27 negative and 21 positive files, we had to eliminate 12 random negative and 6 random positive documents.

Obviously, this modified the dataset considerably and did not allow obtaining pure results, making the task easier for the smaller collections supplemented with documents by the same author, and harder for the bigger collections, from which documents had to be eliminated. This imbalance is reflected in the large correlation coefficient value of -0.320 between the polarity classification results and the authorship accuracy results for each author: the collections attributed less accurately were easier to classify in terms of polarity than the collections attributed with a high degree of accuracy, because the former were supplemented with files with native authorship, whereas from the latter some files, most of them by the native author, were eliminated, in order to balance the dataset.

4.5.1 Experiment 3 Results and Analysis

The resulting collection of documents was used to perform the polarity classification experiment similar to one described in Section 3.1. The resulting mean accuracy of 57.67%, according to our expectation, showed a statistically insignificant increase over the randomly grouped baseline result of 56.47%. The accuracies for the 10 authors correlated positively with these from the experiment described in Section 3.1 for the 10 separate author-groups, with the correlation coefficient being 0.727.

However, from Table 4.2 it is obvious that the authorship attribution algorithm worked with various success rates for different authors, but there is also a strong tendency of the algorithm towards selecting a small number of ‘greedy’ classes and assigning most of the documents to them, while leaving the rest with almost no units. This demands a different evaluation framework, which is outside the scope of this work.

In our case there were four authors, id 1, 4, 6 and 10, representing the ‘greedy’ classes: starting with 15 files, each class gained at least 20 at the end. Initial analysis of the results for these classes shows that when performing authorship attribution to aid sentiment analysis, it is these ‘greedy’ groups that should be focused upon and
evaluated, despite the fact that they do not always represent the actual authorship of the documents.

4.6 Conclusions
The experiments described in this section confirmed the hypothesis that the appraisal polarity is expressed in an individual way by different authors. Moreover, the differences are so considerable that in order to investigate the polarity of documents automatically, a subsequent amount of documents by the same author gives more useful information than a much bigger sample of documents written by other authors.

The personalized approach has improved the results of the polarity classification task. This leads to the intuition that any opinion mining task could be improved if considered in terms of idiolect. We applied an authorship attribution algorithm, to test whether, and to what extent, the personalized approach with known authors could be substituted with automatic authorship attribution. A simple authorship attribution algorithm with medium performance proved capable of supplying useful information for polarity classification task, increasing the performance. As expected, authorship attribution imposed limitations on the dataset in terms of its volume and balance, making the subsequent polarity classification results harder to evaluate.

Thus, we conclude that taking into account authorship, whether known or classified automatically, is a useful direction to take in sentiment analysis. However, the former is not particularly realistic provided that the corpus is extracted automatically from the web, and the latter imposes limitations, especially when applied to a small dataset. This is why we consider investigating features of idiolect representing broader groups of authors in our future work, i.e., groups of authors sharing the same occupation.

Our conjecture is that Personal Sense relates to occupation, or profession, in a particularly strong way. Occupation influences the everyday experience of a person, forming a sociolect, common among individuals of the same profession, but differentiating them from those working in another field. Thus, the next step of our research is to find out, in what way and degree occupation actually forms a sociolect in a person’s language use.
Background information is information about individual experience. It reflects what the writer is or has been. One of the most outstanding background distinctions is the professional one. We assume that one's profession is an important former of individual experience, and this is why we consider it useful to apply the Personal Sense notion to the study of professional background. For example, consider a sentence from a blog dataset described later in this chapter.

“Morning Ladies! I have no idea what is going on with our internet at work today, but I can't see the comments on the blog- are you guys having the same problems?”

The blog is addressed to women and is likely to be have been written by a woman. Moreover, after we read “internet at work”, we understand that the work of the author has to do with a computer in an office, so it does not involve physical work but rather intellectual, and leaves some free time for the author to surf the internet.

In some cases indications of professional background are even more clear in text.

“Part of the problem for female soldiers in the Army is the existence of women like one of my roommates, who utterly reinforces many of the negative stereotypes…”

Here, the author describes women in the army, so that we understand this issue is familiar to them. This is not stated explicitly in the text, but we assume that it is likely that the author has a job in the military area. The more text we read by the current author, the more clear it will become as to whether or not this hypothesis is true.
In this Chapter we consider blog posts written by individuals from which to derive the co-occurrence thesauri. The relation among people in our investigation is their profession. Our hypothesis is that profession forms in a strong way the concept structure of authors, that is revealed in their texts. Using the Personal Sense representation technique, we construct personalized thesauri for a number of blogpost authors, and test that authors sharing the same profession have personalized thesauri more similar to each other than the authors belonging to different occupation. We suggest a model to predict professional background of a text author, that can process high volumes of text, based on Personal Sense identification.

5.1 Motivation: a Personal Sense Thesaurus

In a text by an individual, the concepts that he/she uses acquire a Personal Sense. Depending on unique background information underlying their personal language use, the concepts represented in the text will form a unique semantic structure in terms of their meaning, as influenced by the Personal Sense of the words. The acquired structure can then be considered as a personalized thesaurus for each author in the corpus. Representing the unique background experience information as the Personal Sense semantic structure, the personalized thesauri may be used to compare authors with different background information and finally to infer the background conditions and objective characteristics of the authors: age, gender, profession etc.

We follow [Choudhury et al, 2008] in investigating the semantic inter-relatedness of words for different persons, especially for authors with different occupations. However, while they define every profession in terms of its semantic relatedness to every other notion - suggesting a manually created ontology of the semantic field and establishing an ontology-based semantic relatedness score (see Table 5.1 for examples) - our consideration is to derive the semantic relatedness of the notions from text, and thus infer the author’s profession.
Table 5.1. Semantic relatedness scores for some pairs of Motion Picture Industry concepts

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
<th>Total Semantic Relatedness Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method Acting</td>
<td>Actor</td>
<td>3</td>
</tr>
<tr>
<td>Cross-cutting</td>
<td>Editor</td>
<td>5</td>
</tr>
<tr>
<td>Cross-cutting</td>
<td>Actor</td>
<td>119</td>
</tr>
<tr>
<td>Motion Picture</td>
<td>Actor</td>
<td>4</td>
</tr>
<tr>
<td>Motion Picture</td>
<td>Editor</td>
<td>4</td>
</tr>
</tbody>
</table>

We use the algorithm described in [Yoshida et al, 2003] constructed ‘to find relations among people based on their interests and knowledge’. We apply it to measure the similarity between the personal thesauri.

5.2 The Dataset: the Professional Background Information

We work with the blog text corpus described in [Schler et al, 2006]. It includes around 71,000 blogs from blogger.com obtained in August 2004, containing at least 200 common English words and information about the authors' age and gender, and profession in most of the cases.

For the initial experiment we investigated 10 randomly-selected blogs annotated as ‘Accounting’ and 10 annotated as ‘Military’, with the annotation referring to the author's self-annotation in terms of their occupation or profession. We use a Natural Language Toolkit (NLTK)\(^6\) for the Python programming language for most of the text analysis. All the texts were annotated automatically with parts-of-speech using the Maxent Treebank POS-Tagger\(^7\) built-in into the NLTK. Some POS-tag examples

---

\(^6\) Available at http://www.nltk.org/.

\(^7\) Part-of-speech tagging (POS-tagging) is a procedure of assigning a part of speech to every word in a text. Knowing the part-of-speech for every word in a text is essential for most of the tasks in Natural Language Processing, e.g. Word Sense Disambiguation, Machine Translation etc. Tagging used to be performed or checked manually ([Francis et al, 1964]) before computer processing started to play an important role in linguistic analysis. However, as manual part-of-speech tagging is very time-consuming, now there is a number of algorithms used for POS-tagging. The standard methods of POS assignment require a pre-tagged text and use it as a learning sample to train statistical models. Then the resulting models are applied to the text in question. The underlying principle of all the algorithms is that a tag is selected that has the highest probability for the given word in the given circumstances (see [DeRose, 1990]). For example, the Hidden Markov Models –
used in this chapter are presented and described in Table 5.2 ([Treebank Project, 2010]).

The goal of the experiment is as follows. Firstly, to represent Personal Sense of words in text by different authors using co-occurrence distributional word meaning techniques. Secondly, to form personalized thesauri based on the inter-relations between words. Thirdly, to infer relations among authors’ in terms of their background, particularly their occupation, from the distance measures between their personalized thesauri. Thus, we continue exploiting the interrelation of private states expressed in text and authorship attribution, described in the previous section.

Table 5.2. Examples and description of the Treebank part-of-speech tagset.

<table>
<thead>
<tr>
<th>POS tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>determiner</td>
<td>the</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/subordinating conjunction</td>
<td>in, like</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>many</td>
</tr>
<tr>
<td>NN</td>
<td>noun</td>
<td>part, problem, soldiers</td>
</tr>
<tr>
<td>PR</td>
<td>pronoun</td>
<td>I</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>clearly, not</td>
</tr>
<tr>
<td>VB</td>
<td>verb</td>
<td>live, has</td>
</tr>
</tbody>
</table>

5.3 Experiment: Construction of Personalized Thesauri Representing Authors’ Professional Background

First, a representative group of words was selected to investigate their Personal Sense. In our preliminary experiments these were the 10 most-frequent meaningful nouns occurring in each blog. The results indicated that the meaning of these nouns based tagger selects the most probable sequence of POS tags in given circumstances. The Brill tagger ([Brill, 1994]) operates by assigning a default tag to every word and applying a number of rules in a number of cycles, gradually achieving considerable accuracy. The algorithm used in our application is the Maximum Entropy algorithm ([Ratnaparkhi, 1996]) based on selecting a tag with the highest probability for the given word against a number of initial circumstances, trained on the Penn Treebank corpus ([Treebank Project, 2010]).
is so common that it is hard to investigate their automatically-derived Personal Sense and compare it to their actual use in texts.

After manual investigation another group of nouns was selected: \{career, power, war, law, rule, army, man, woman\}. These were chosen because after manual text analysis their Personal Sense appears to contain different relations of the people to the concepts underlying the words. We also added the 25 most frequent nouns to the dataset: \{dad, mom, person, company, movie, husband, test, parent, dinner, world, head, place, child, thing, school, way, life, family, house, work, job, car, home, girl, friend\}. Thus, an extended group of nouns was selected for the experiment, consisting of 33 nouns, including both the 25 most frequent nouns and the 8 manually selected nouns.

Next, their Personal Sense was represented. Each word was represented by its co-occurrence distributional vector. Every dimension is a word occurring in the context(s) of the investigated nouns. As we used POS-tagged texts, by ‘word’ here we mean a pair of \{word, POS-tag\}. As an example, the contexts of the target word ‘woman’ taken from the following sentences are presented:

- “Part of the problem for female soldiers in the Army is the existence of women like one of my roommates, who utterly reinforces many of the negative stereotypes...she’s whiny, she runs to the doctor every time she sneezes, she’s lazy and does everything she can to get out of work and physical training.” (ex. 5.1)

- “SGT B, however, the woman I live with, has clearly not had many female friends.” (ex. 5.2)

We observed 4 different definitions of context, presented in Table 5.3.
Table 5.3. Examples for 4 context definitions.

<table>
<thead>
<tr>
<th>Number of context</th>
<th>Description of the context of the word ( w_1 )</th>
<th>Examples of the contexts for sentences in ex. 5.1 and ex. 5.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A set of pairs of words; every pair consists of a word immediately to the left of ( w_1 ) and a word immediately to the right of ( w_1 )</td>
<td>(of/IN, like/IN) (the/DT, I/PR)</td>
</tr>
<tr>
<td>2</td>
<td>A set of the same words as in (1), not organized in pairs, but for each word it is indicated if it belongs to the left or to the right context</td>
<td>l_of/IN, r_like/IN l_the/DT, r_I/PR</td>
</tr>
<tr>
<td>3</td>
<td>The same set of words as in (2), but no left- or right-side context indication.</td>
<td>of/IN, like/IN the/DT, I/PR</td>
</tr>
<tr>
<td>4</td>
<td>All the words belonging to the sentences in which ( w_1 ) occurs, except ( w_1 ).</td>
<td>Part/NN of/IN the/DT problem/NN for/IN female/JJ soldiers/NN in/IN the/DT Army/NN SGT/NN B/NN , however/RB , has/VB clearly/RB</td>
</tr>
</tbody>
</table>

The context window definition of only one word to the left and one to the right of the target word is very narrow, and we expect that it might not yield the best results. However, it is this type of context that we investigate in the current work for two reasons: first of all, using this context definition is very helpful in manual examination of the resulting Personal Sense structures. Moreover, together with the context option number 4, which takes into account all the words occurring in the same sentence with the target word, these context definitions present the opposite extremes of a very broad and a very narrow context, this fact giving important insights, as we expect, for analyzing the final results.

Every word in context acquires a weight of the conditional probability of the word given the target word \( w_1 \) (see formula 5.1).

\[
P(w \mid w_1) = \frac{P(w \cap w_1)}{P(w_1)} = \frac{\text{Frequency}(w, w_1) \text{co-occurring}}{\text{Frequency}(w_1)}
\]

(f. 5.1)
We applied clustering to the obtained vector space to find out which understanding of context yields a vector space more adequate to our manually derived expectations. We used The Group Average Agglomerative clustering technique [Romesburg, 1984] implemented in NLTK to represent structure of words. The algorithm starts with each word as a separate cluster, merging two clusters into one on each step. Thus the words lying closer to each other are merged into one cluster on earlier steps, whereas the words joining them on later steps represent similarity less significant. For clustering and computing similarity between two words in a text we used the cosine similarity measure between the target words (see the Cosine for Conditional Probabilities formula, (f. 5.2)).

\[
\text{CosCP}(w_1, w_2) = \frac{\sum_{w \in \text{cont}(w_1) \cup \text{cont}(w_2)} (P(w \mid w_1) \cdot P(w \mid w_2))}{\sqrt{\sum_{w \in \text{cont}(w_1)} P(w \mid w_1)^2} \cdot \sqrt{\sum_{w \in \text{cont}(w_2)} P(w \mid w_2)^2}}
\]

(f. 5.2)

The final stage experiments are dedicated to comparing the obtained personalized thesauri with the techniques similar to that described in [Yoshida et al, 2003]. We compared three different thesauri distance measures. For each of them we used the basic formula presented in [Yoshida et al, 2003].

\[
d = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} (d_{i,j}')^2}
\]

(f. 5.3)

\[
d'_{vw} = \begin{cases} 
sim^S_v - sim^T_w, & (v, w) \in S, T \\
\min(sim^S_v, sim^T_w), & (v, w) \in S, \notin T \\
\min(sim^S_v, sim^T_w), & (v, w) \in T, \notin S \\
x, & (v, w) \notin S, \notin T 
\end{cases}
\]

(f. 5.4)

The first distance measure, \(d_1\), was exactly the same as described in [Yoshida et al, 2003]. That means, in (f. 5.3) \(q\) is equal to \(m^2\), the squared number of words in the target words vocabulary; and \(x\) in (f. 5.4) is assigned the value 1.
For the second measure, $d_2$, $x$ in (f. 5.4) is equal to 0. In other words, when the word-pair $(v, w)$ is not present in neither of the two thesauri $S$ and $T$, we add nothing to the sum.

For $d_3$, $x$ is also equal to 0. But $q$ in (f. 5.3) represents the number of the word pairs present in at least one of the two vocabularies $S$ and $T$.

The difference between the thesauri distance formulae 5.1, 5.2 and 5.3 lies in approaching the words absent from the personal blog vocabulary in a different way. We expect that this difference will affect the results in a considerable way, because the two groups of blog authors that we are investigating manifest differences in terms of their vocabulary usage: the ‘Military’ authors tend not to use all of the target words in their blogs, whereas for the ‘Accounting’ authors this is usually not the case.

We expect that at least one notion of ‘context’ constitutes the Personal Senses of the observed words so that the Personal Sense and their inter-relations represent the background and particularly occupational differences between the authors to a sufficient degree, and the resulting average distances between personalized thesauri are bigger for authors having different occupation than for those sharing one.

### 5.4 The Experiment Results: Comparing the Personalized Thesauri by Authors with Different Occupation

Results obtained by clustering the co-occurrence distribution vectors show that the co-occurrence vector space confirms the manually derived suggestions for two ‘opposite’ definitions of context: the ordered pair of one word on the left and one on the right(1), and all the words in the whole sentence containing the target word(4). The former definition captures Personal Sense relations for words with higher frequency in the text; whereas the latter naturally underlines relations between words occurring with a very low frequency in a text and occurring together in the same sentence. We analyzed pairs of words acquired on the first clustering step; for the former context definition there are 24 such pairs for 20 blog texts; more than 70% of them are in some way accountable in terms of Personal Sense. For example, in the case of an author who dedicated an considerable part of her blog to discussing a special role of women in army, the first-step pair acquired was ['army', 'woman'], and
another author, who also discussed war, but from the perspective of military laws and government, the first-step result was ['war', 'rule'].

It is not the goal of our investigation to give a thorough analysis of the clustering results. However, clustering is very important as an intermediate step for demonstrating that the Personal Senses cause the target words to represent different structures for different blog authors and thereby establishing a basis for the next stage experiment.

The personalized thesauri comparison results for the 33 target words, for 4 different types of context and the 3 distance functions are presented in Table 5.4.

Our assumption was that the personal word sense interrelation structures would be more similar among authors belonging to the same occupational background, than between the authors belonging to different professional groups. We represented the Personal Sense interrelations of each author as a personalized thesauri based on their texts. The distances between the thesauri were computed for each pair of authors. We divided the pairs of thesauri into 3 types depending on their authors:

- (1) (T[military], T[military])
- (2) (T[accounting], T[accounting])
- (3) (T[military], T[accounting]), and (T[accounting], T[military]),

where T[x] represents the thesaurus of an author belonging to the occupation x.
Table 5.4. The personalized thesauri comparison results for blogs by authors of the same and different occupation

<table>
<thead>
<tr>
<th></th>
<th>Thesauri distance comparison between</th>
<th>Statistical significance according to the paired t-test (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>d₁</td>
</tr>
<tr>
<td>Context:</td>
<td>‘Military’ vs different</td>
<td>‘Accounting’ vs different</td>
</tr>
<tr>
<td>1</td>
<td>&gt;</td>
<td>&lt;</td>
</tr>
<tr>
<td></td>
<td>0.74</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>&gt;</td>
<td>&lt;</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>&gt;</td>
<td>&lt;</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>&gt;</td>
<td>&lt;</td>
</tr>
<tr>
<td></td>
<td>0.47</td>
<td>0.68</td>
</tr>
<tr>
<td>d₂</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>&lt;</td>
<td>&gt;</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>&lt;</td>
<td>&gt;</td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>&lt;</td>
<td>&gt;</td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td></td>
<td>0.61</td>
<td>0.59</td>
</tr>
<tr>
<td>d₃</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>&lt;</td>
<td>&gt;</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>&lt;</td>
<td>&gt;</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>0.99</td>
</tr>
<tr>
<td>3</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td></td>
<td>0.56</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.7</td>
</tr>
</tbody>
</table>

We have computed the distances between the thesauri in groups 1, 2 and 3 separately. According to our assumption, the distances in groups (1) and (2), i.e. between thesauri by authors of the same occupation, would be smaller than for the distances of group (3), representing the distances between the different occupational group thesauri. An example of the distances between the personalized thesauri is presented in Table 5.5.

In spite of the fact that the results did not unanimously confirm our overall hypothesis, they indicate some very important tendencies and provide evidence...
supporting our contention that the Personal Sense technique is useful for discriminating between authors’ occupations.

**Table 5.5.** Examples of distances for some pairs of the personalized thesauri

<table>
<thead>
<tr>
<th>Author 1 id</th>
<th>Author 2 id</th>
<th>Distance score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1198592.fem.Accounting</td>
<td>1334509.fem.Accounting</td>
<td>0.149</td>
</tr>
<tr>
<td>928859.fem.Military</td>
<td>3331865.mal.Military</td>
<td>0.187</td>
</tr>
<tr>
<td>1198592.fem.Accounting</td>
<td>928859.fem.Military</td>
<td>0.220</td>
</tr>
</tbody>
</table>

The results with high statistical significance (p-value < 0.01), obtained for the context definition 1, for distance measures 2 and 3, for the ‘Military’ occupation group, confirm the hypothesis and show that the inter-group distances are lower than the intra-group distances. This could also indicate the fact that the military group is easier to discriminate than the accounting group, as the personalized thesauri for these group are closely related to each other, whereas for the accounting group the structure is sparser. In most of the cases little statistical significance was achieved, indicating that more target words and more texts should be analyzed.

It is even more important to note that we have obtained contradictory results for different distance options, according to our expectations described earlier in this section. In most of the cases distance 1 yields opposing results to those obtained using distances 2 and 3, in terms of the inter- and intra-group distances between the occupations. These formulas mainly differ in the way that they handle words that did not occur in the texts by specific authors. As the results for the different distance formulas suggest, the difference in approaching these words alters the results dramatically, regardless all the other options. This means that for the current data and context definition, the absence or presence of words in the authors’ vocabularies outperform the Personal Sense differences computed for the words. This was our expectation that was confirmed by using the words that were not contained in all of the texts. Although finally it is important to combine various types of information for the occupation classification, at the current stage we find it useful to investigate the Personal Sense structures in isolation.

The only context definition that yielded more consistent results for different distance functions is the context number 4, containing all the words in the sentence where the target word appears. This confirms a consideration that the broader the context
definition is, the more it influences the results, regardless of the different distance measures. However, the very low statistical significance in the case of such context shows that the number of target words should be increased, and that it would of benefit to apply a filtering technique to Personal Sense representations in the thesauri.

5.5 Conclusions

Our experiments confirm that for target words of certain number and frequency, for the most fine-grained context definition implying ordered word-pairs of the words immediately to the right and to the left of the target word, and the distance measure for thesauri comparison sensitive to presence or absence of the target words in the blog, the differences between the resulting thesauri can effectively represent the differences among the occupation background of the authors of texts. Personalized concept structures were constructed from texts by different authors, taking into account the Personal Sense of the words they used. The resulting structures were unique for every blog writer, representing the unique Personal Sense. The personal concept structures have been utilized to infer the authors’ perspective from their writings. The personalized thesauri for writers belonging to different professions were compared using a thesaurus similarity scale. The results prove that different occupations are not equally easy to analyze, i.e., some of the occupations present similarities in the personalized thesauri that can be easily detected, whereas others require additional experiments with more refined options. The results confirmed that within certain limitations, the method of representing the similarities in the personalized thesauri could be used to reflect similarities in occupation of the authors.
Chapter 6: Identifying Subjective Statements in News Titles Using a Personal Sense Annotation Framework

Subjective language identification is a very important direction in opinion mining. Subjective language is language containing information about private states ([Wiebe et al., 2004]): opinions and emotions. The goal of subjective language identification is to identify that a private state is expressed, without going into detail about its polarity or its specific emotion. On one hand, it is a preliminary stage in opinion mining: before identifying an opinion as positive or negative, it is necessary to identify it as an opinion, as opposed to a fact description etc. Furthermore, it may serve as a technique for separating facts from points of view, classifying opinionated text and identifying ideological perspective of the author.

Consider the following examples, taken from a debate about establishing a border between Israel and Palestine

*The “green line” is invisible, undocumented and unfounded in international law[...] it sets a precedent of substituting principles of international law with agreements signed under duress.* (ex. 6.1)

*Despite these trans-boundary movements, the line remained an important point of separation between the two territories[...] the green line—with some minor deviations—has the greatest likelihood of constituting the formal international boundary between two independent states.* (ex. 6.2)

Both pieces contain information about the green line, not serving as a border between two independent states yet - and this is where opinions begin - the first
author believes it to be illegitimate and gives a negative assessment of the possibility of it becoming a formal and legal object. The second author, on the other hand, assesses it positively as one of the formers of two independent states.

Both authors describe the same phenomena, but use different words relating to it. The words 'undocumented, invisible, duress' in the first passage and 'important, independent' in the second are the clues that help us detect subjectivity expressed. An automatic subjectivity identification tool uses broadly the same technique: it captures subjective clues in text and relates them to certain objects or topics of discussion.

In this chapter, we provide an annotation schema for the Personal Sense ‘target’ and ‘indicator’ constructions covering emotion, polarity and subjectivity-objectivity in terms of the Personal Sense. We proceed to analyze the subjectivity-objectivity issue. Assuming that the subjective Personal Sense patterns are constructed in text on a regular basis using lexical and syntactic elements, we perform an experiment on the automatic detection of subjectivity in text, as opposed to objective expressions not containing any subjective emotion. First, we demonstrate with a preliminary experiment that subjective expressions are more accurately described using a combination of lexical and syntactic information than by using lexical means only. Next, we select a number of lexical and syntactic features for the identification of the subjective patterns in text. We apply the Personal Sense technique to pairs of words, at least one of which is a noun: thereby identifying the Personal Sense of the noun in the pair. We argue that the suggested features, including the syntactic path between the words in the pair and lexical information about the Personal Sense indicator word, are useful for the identification of the subjective Personal Sense. We use the resulting subjective and objective word-pairs for the subjectivity classification applying the suggested feature set. Thus we learn to identify automatically the word-pairs, connected by a certain syntactic path, bearing an emotion, as opposed to the pairs that do not bear any emotional content. The results confirm our expectations and demonstrate the lexico-syntactic features to be useful for the identification of subjectivity indicators and the targets which receive the subjective Personal Sense.

6.1 Preamble: Intentional Object in Subjectivity Analysis

There has been done considerable work on the identification of private states in language [Pang et al, 2008]. However, it is important to underline the meaning of the
emotions intentional object, which has not yet been investigated from a lexicosyntactic viewpoint. The authors of [Scherer, 1999] discuss intentionality and the intentional object as an inherent characteristic of emotions, as well as appraisal. While the focus of previous work has been on the overall emotion, mood, or more broadly, private state of a sentence or even a text, we pursue a more fine-grained goal of identifying emotions intended at specific objects, attaching a subjective polar or emotional Personal Sense to the object in the text.

Emotion is focused on an object or event relevant to a person's motivation ([Scherer, 1999]). This stimulus is called an intentional object of the emotion. A word in a text acquires a subjective emotional Personal Sense if it is the intentional object of an emotion expressed in the text. Thus, there should be the object of the emotion, presented by a word in the text: we call this the target word and investigate its Personal Sense; and a word or a construction that indicates the emotion (i.e., adds the emotional Personal Sense to the target word) is called the indicator. In practice, there is another important element that influences the polarity and emotion of the target together with indicator: the intermediate element that may alter or increase the target polarity and emotion against the indicator ones. Some examples of the obtained Personal Sense structures will be presented later in this Chapter, in Table 6.2.

6.2 A News Headlines Dataset Annotated with Emotions and Polarity
The dataset that we are considering in this section has been described in [Strapparava et al, 2007] and is widely used in research ([Bhowmick et al, 2009], [Bao et al 2009]). It consists of 1250 news headlines from major newspapers as BBC News, and from the Google News search engine. The titles are annotated in a fine-grained manner with six basic emotions (Anger, Disgust, Fear, Joy, Sadness, Surprise) and valence (Positive/Negative). For each emotion there is a scale ranging from 0 to 100, indicating the degree of the emotion presence in the sentence. Valence is represented by a number ranging from -100 to +100, with 0 indicating a neutral headline, -100 and +100 represent a highly negative and a highly positive headline respectively.

6.3 Motivation: a New Fine-Grained Annotation Based on Personal Sense
In order to demonstrate the motivation for identifying Personal Sense in headlines, let us consider an example.
This sentence is characterized, according to the provided gold standard, by positive (31 out of 100) valence and a considerable contribution of joy, surprise, fear, and a slight impact of sadness and anger (in descending order). Consider an artificial counterexample, not occurring in the current dataset:

'The hostage supposed to be freed is dead' (ex. 6.4)

According to our intuition, the sentence should acquire negative valence, and the dominant emotion (the emotion characterized by the highest impact number) should be sadness. However, if we analyze the sentence as a whole in terms of emotional words contained in it, we will get the same pair 'freed', 'dead' for both sentences, and an additional 'feared' for the first one. This does not give us insight about the opposing polarity values for the examples. It is only when we approach the syntactic level and realize that the predicates of the two sentences contain the opposite meaning in relation to the same passive subject, that we can understand why the two examples containing the same meaningful set of words acquire opposite polarity value and different dominant emotions.

The subject of the sentence also plays a very important role in the analysis. In example 6.3, two opposing emotions are expressed about the ‘hostage’. The joy about the hostage being free overweighs the fear about him/her being dead and becomes an overall sentence emotion; but the fear about the hostage is also present. It is only when we refer both of these emotions to their intentional object – the hostage – that we can compare them in terms of their impact on the overall sentence emotion, and conclude that the joy overweighs the fear, being expressed in a higher-level syntactic dependency, i.e. the predicate of the sentence.

For these reasons, we consider it very important to analyze sentiment and emotion in a fine-grained way, attributing emotions to the Personal Sense of their intentional objects, with syntactic paths serving as formal representation of such attribution (compare the usage of linguistic clues in [Wiebe et al, 2004]).

Of particular interest to us is the fact that the intentional object of an emotion acquires emotional Personal Sense. In example 6.3, the Personal Sense of ‘hostage’ contains the emotions of fear and joy, expressed by ‘feared (dead)’ and ‘is freed’.
Moreover, the fact that the author expresses fear about the hostage being dead, and happiness about them being freed, i.e. the negative Personal Sense of the word 'hostage' expressed by 'feared (dead)', and the positive one, represented by 'is freed', delivers some important information about the author: the author clearly opposes the actions of the kidnappers. Although this seems normally the case, it could be different if, for example, the author wrote an extremist slogan and argued for the same demands as the group of terrorists or kidnappers. Thus, Personal Sense indicates social affiliation of the author, and depends largely on the intended reader: the customer who buys newspapers that describe ideas in an appropriate manner. Among other characteristics, the author expresses appropriate emotions in the Personal Sense of the objects and events described.

6.4 The Personal Sense Annotation Schema: Indicator, Target, and Intermediate Elements

We follow the theoretical considerations that emotions are directed at objects or events causing them, named intentional objects and appraised in terms of personal motivation [Scherer, 1999], and provide the new annotation framework for Personal Sense in text.

An emotion expressed in a sentence consists at least of a pair of words [indicator, target], where indicator expresses the presence of emotion itself, and target denotes the intentional object of the emotion.

We will exemplify this with the familiar sentence from example 6.3. The author explicitly expresses fear about the hostage being dead, and implicitly introduces happiness about him/her having been freed. The fear about the hostage is due to the possibility of his/her death, thus in the first case, the Personal Sense indicator is the word ‘dead’. ‘Feared’ in this case serves as an intermediate element, which is described further. The word ‘dead’ clearly indicates negative polarity, but it can ascribe different Personal Sense emotion to its target word: ‘fear’ in the case of example 6.3, but ‘sadness’ in the case of example 6.4. The specific emotion does not depend always on the indicator word, but on the syntactic connection between the indicator and the target, and the nature of the target word itself in some cases. We cannot identify clearly if the word ‘dead’ in an isolated position implies either ‘sadness’ or ‘fear’. Accordingly we do not attribute the implied emotion to the indicator word, but rather to the target word. On the other hand, polarity is attributed
to both of them, as it is usually unambiguously represented by the indicator word (clearly negative in the case of ‘dead’), and is clearly present in the Personal Sense of the target word (we feel bad about the hostage fearing that he/she is dead), but can acquire the opposite value in some cases. Consider the following example:

‘Rights group halts violent Nepal strikes.’

where the (violent) ‘strikes’ contain negative polarity, but ascribe a positive Personal Sense to the word ‘group’, because it ‘halts’ the strikes.

This example demonstrates that in some cases an additional element is necessary in order to describe the Personal Sense relation correctly. As described in the example above, it is the word ‘halt’. It transforms the target polarity to the opposite of the indicator one, causing the ‘group’ to acquire a positive Personal Sense, despite the negative polarity of the ‘strikes’ indicator. Because of this transformation it is useful to include this word into the annotation as an ‘intermediate’ element. Intuitively it is a word that stands between the indicator and the target in terms of syntax, and it can transform the polarity of the relation radically, as in the example sentence above. However, in the dataset we encountered a significant number of such words that clearly occupy the same syntactic position, intermediate between the indicator and the target, but which do not transform the polarity, although they can be substituted by words that would transform it. Consider the following examples:

‘Snow causes airport closures in Britain.’

‘Stenson defends his title at Dubai.’

If we substitute ‘defends’ by ‘loses’ in ex 6.7, we will get a negative Personal Sense for ‘Stenson’, despite the positive ‘title’. The word ‘defends’ in this case occupies a very important intermediate position, but does not bring a transformation to the polarity. This is why we also consider such words as intermediate elements, and define a modality attribute for them, which takes the negative value when the intermediate element in question changes the polarity between the corresponding indicator and target elements, and the neutral or positive value when polarity value is preserved. The resulting annotation schemas are presented in Table 6.1.

The table describes and gives examples of the indicator, intermediate and target annotation schemas designed for our annotation. The indices in the ‘Attribute value’
column are used to highlight the attributes that ought to have the same or corresponding value:

- id of the indicator, intermediate and target should be the same, in order to process the words in a single Personal Sense relation;

- polarity and emotion of the target should correspond to each other:
  - joy should be used with positive polarity,
  - the rest of the emotions with negative polarity;
  - surprise can occur with both.

Table 6.1. Annotation schemas for the Personal Sense annotation of polarity and emotion in ex 6.5.

<table>
<thead>
<tr>
<th>Element name</th>
<th>Attribute name</th>
<th>Attribute type, possible values</th>
<th>Attribute value</th>
<th>Attribute value for ex 6.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicator</td>
<td>id</td>
<td>integer</td>
<td>id,</td>
<td>67</td>
</tr>
<tr>
<td>'violent'</td>
<td>polarity</td>
<td>string: neg, pos</td>
<td>pol,</td>
<td>neg</td>
</tr>
<tr>
<td>Target</td>
<td>id</td>
<td>integer</td>
<td>id,</td>
<td>67</td>
</tr>
<tr>
<td>'group'</td>
<td>polarity</td>
<td>string: neg, pos</td>
<td>pol,</td>
<td>pos</td>
</tr>
<tr>
<td></td>
<td>emotion</td>
<td>string: anger, disgust, fear, joy, sadness, surprise</td>
<td>emotion,</td>
<td>joy</td>
</tr>
<tr>
<td>Intermediate</td>
<td>id</td>
<td>Integer</td>
<td>id,</td>
<td>67</td>
</tr>
<tr>
<td>'halts'</td>
<td>modality</td>
<td>String: negative, neutral, positive</td>
<td>mod,</td>
<td>negative</td>
</tr>
</tbody>
</table>

Table 6.2 contains some actual examples of the polarity and emotional Personal Sense relation from the dataset.

The headlines are not annotated explicitly with subjectivity/objectivity: an element is supposed to acquire a subjective Personal Sense if and only if it is annotated with any emotion and any polarity. Any word serving as a target would thus acquire subjectivity, and its indicator would at the same time be a subjectivity indicator. Thus we get a hierarchy of classes: first, a word-pair ([group, violent] in ex. 6.5) belongs to the subjective class, if it acquires any emotion and polarity, and the objective class, if it does not acquire any. Next, if the pair is subjective, the indicator
and the target have a positive or a negative polarity. Polarity may be different for the indicator and the target (as in the case of ex. 6.5: positive for group, negative for violent), and then an intermediate element is introduced that explains the difference (halts in ex. 6.5). Moreover, the target element having negative polarity acquires Personal Sense containing an emotion: anger, disgust, fear, sadness, or surprise. The target with positive polarity acquires joy or surprise. (In the case of ex. 6.5 the emotion attached to group is joy.)

Table 6.2. A sample of the annotated Personal Sense relation components.

<table>
<thead>
<tr>
<th>Id</th>
<th>Sentence</th>
<th>Emotional relation elements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Indicator</td>
</tr>
<tr>
<td>1</td>
<td>Mortar assault leaves at least 18 dead.</td>
<td>Dead</td>
</tr>
<tr>
<td>2</td>
<td>Mortar assault leaves at least 18 dead.</td>
<td>Dead</td>
</tr>
<tr>
<td>3</td>
<td>Goal delight for Sheva.</td>
<td>Delight</td>
</tr>
<tr>
<td>4</td>
<td>Goal delight for Sheva.</td>
<td>Delight</td>
</tr>
<tr>
<td>5</td>
<td>Nigeria hostage feared dead is freed.</td>
<td>Dead</td>
</tr>
<tr>
<td>6</td>
<td>Nigeria hostage feared dead is freed.</td>
<td>Freed</td>
</tr>
<tr>
<td>7</td>
<td>Bombers kill shoppers.</td>
<td>Kill</td>
</tr>
<tr>
<td>8</td>
<td>Bombers kill shoppers.</td>
<td>Kill</td>
</tr>
<tr>
<td>9</td>
<td>Vegetables, not fruit, slow brain decline.</td>
<td>Decline</td>
</tr>
<tr>
<td>10</td>
<td>Rights group halts violent Nepal strikes.</td>
<td>Violent</td>
</tr>
<tr>
<td>11</td>
<td>Rights group halts violent Nepal strikes.</td>
<td>Strikes</td>
</tr>
<tr>
<td>12</td>
<td>Snow causes airport closures in Britain.</td>
<td>Closures</td>
</tr>
<tr>
<td>13</td>
<td>Stenson defends his title at Dubai.</td>
<td>Title</td>
</tr>
</tbody>
</table>

By annotating all the cases of pairs of subjective (or emotional) expressions and their intended objects, we have defined the area of “expressing subjectivity/emotion towards an object in text” extensionally. In other words, we have covered all the actual occurrences of the emotional expressions towards an object in the current dataset. The emotional expression towards an object can be considered a semantic relation, similar to date of birth or headquarters described in [Suchanek et al, 2006]. It is more complex in our case, in the sense that it covers a variety of linguistic phenomena from a formal point of view. It cannot be described by a single syntactic construction or co-occurring with items belonging to a single lexical class. In the
following section we consider the emotional relation with the formal lexical and syntactic phenomena in text.

6.5 The Preliminary Experiment: Lexical and Lexico-Syntactic Approaches in Personal Sense Definition

Lexical items are one of the traditional features used to define semantic relations in text [Gamallo et al, 2005]. Our first assumption was that the emotional relation is solely characterized by the indicator lexical items, i.e., given a specific word, we assume that it does or does not indicate a subjective expression. To test this assumption, we computed the conditional probability for a lexical item to indicate a subjective expression, given that a lexical item is a subjective clue. The mean result was 83.64%, with the standard deviation of 0.26. This is a modest result, indicating the highest possible subjectivity identification accuracy that we can get, if we correctly identify a word to be subjective or objective. The main reason that we see for such a modest estimation is that most of the words can indicate subjectivity in one case, while preserving an objective meaning in another case; in other words,

“many expressions with subjective usages have objective usages as well”

(citation 6.1, [Wiebe et al, 2004]). We exemplify this consideration with the actual dataset, even with such a seemingly non-ambiguous (in terms of subjectivity and emotion) word as ‘good’.

Consider the following two sentences:

‘PM: Havana deal a good experiment.’ \hspace{1cm} (ex. 6.8)

and

‘Bad reasons to be good.’ \hspace{1cm} (ex. 6.9)

In both sentences the word ‘good’ is automatically tagged as an adjective. In example 6.8 it is annotated as a subjectivity indicator, with the target word being ‘experiment’, but it does not indicate any subjectivity in example 6.9 according to our annotation, as there is no obvious target word or intentional object present in the sentence. Thus, the lexical information alone does not define the subjectivity area accurately enough.
We assume that an additional feature useful for the subjectivity-objectivity distinction is the syntactic dependency path connecting the indicator and its target word. After we added this type of information, we computed the conditional probability of an expression being subjective, given a subjective lexical item and a subjective dependency path. This time the mean result was 92.50% with the standard deviation of 0.18, which proved to be higher than the lexical-only based result with a 99% statistical significance.

We conclude that emotional Personal Sense may be characterized more accurately by a combination of lexical information and syntactic information. To develop on the statement from citation 6.1, if “expression” is considered not a lexical item alone, but a pair consisting of the lexical item and a syntactic path to the potential target, then the expressions with subjective usages are more likely to have no objective usages. To classify whether or not a word indicates emotional Personal Sense in a sentence, it is useful to identify the dependency construction that contains the lexical item, because the same lexical items vary in their subjective and emotional impact depending on the dependency structures in which they occur.

6.6 The Experiment: Identifying Subjective Personal Sense

In the current chapter we are investigating the Personal Sense of nouns, although they are not the only part of speech providing the targets for the Personal Sense annotation. With the targets of the Personal Sense relation limited to nouns, 475 ‘target-indicator’ word pairs were annotated and automatically extracted.

Our assumption is that the new framework for subjectivity classification based on the Personal Sense detection yields considerable results, even when applied to fine-grained subjectivity recognition. The goal of the experiment is to test the performance of the lexico-syntactic technique based on our Personal Sense framework, and to estimate the value of the suggested lexical and syntactic features for the subjectivity classification.

We have performed a classification experiment with pairs of words, the first being a noun, and the second one potentially representing the Personal Sense indicator. There were 475 word-pairs annotated with subjectivity, and 17,500 pairs containing no subjectivity. We divided the neutral items into 37 random groups and performed the classification experiment with each of them, in order for the dataset to be
balanced with respect to subjective and objective newstitles, i.e. there were the same number of pairs representing appraisal and emotional Personal Sense, as the number of pairs containing no Personal Sense, in each classification subset.

### 6.6.1 Lexical and Syntactic Classification Features

We investigated a number of features for the classification of pairs of words as containing or not containing subjectivity. First of all, this was the word itself, the lexical item (LEX) that was potentially a Personal Sense indicator for the noun in question; and the part of speech for the lexical item (POS). Secondly, this was the syntactic path (PATH) from the potential indicator to the potential target noun, according to the dependencies identified by the Stanford parser. As in real-life classification we would often encounter new lexical items as potential Personal Sense indicators, accordingly we used some features derived from SentiWordNet ([Esuli et al, 2006](http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm)) and General Inquirer in order to represent the degree of subjectivity for lexical items themselves.

SentiWordNet contains information about a number of senses for each word (115,400 senses altogether), often more than one, sometimes more than ten or twenty. Each of the senses is characterized by a synset and a score for positivity and negativity, both ranging from 0 to 1. For every word we used the following 8 features derived from its SentiWN profile. The features are presented in Table 6.3.

<table>
<thead>
<tr>
<th>Name of the feature</th>
<th>Significance of the feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMALL</td>
<td>number of the word's SentiWN senses</td>
</tr>
<tr>
<td>BIGGESTVALUE</td>
<td>maximum ‘positivity’ or ‘negativity’ value with the respective sign</td>
</tr>
<tr>
<td>BIGGESTSUM</td>
<td>he biggest sum of the positive and the negative value with the respective sign</td>
</tr>
<tr>
<td>NUMPOSSUM</td>
<td>the number of senses for the current word for which this sum is positive</td>
</tr>
<tr>
<td>NUMNEGSUM</td>
<td>the number of senses for the current word for which this sum is negative</td>
</tr>
<tr>
<td>NUMBIGPOSSUM</td>
<td>number of senses for which the sum was higher than 0.25 and positive</td>
</tr>
<tr>
<td>NUMBIGNEGSUM</td>
<td>number of senses for which the sum was higher than 0.25 and negative</td>
</tr>
<tr>
<td>BIGGESTNUMOTHE</td>
<td>maximum positive or negative value in the sense, for which the other value</td>
</tr>
</tbody>
</table>

8 Available at http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm.
The General Inquirer is a tool for content-analysis of textual data, based on a number of word categories, whose distributions are calculated for a text and used for different goals, with polarity classification and emotion recognition among them. We used 14 categories from the General Inquirer vocabulary, namely: Neg, Ngtv, Virtue, EMOT, Pstv, Pos, Hostile, Pleasure, Vice, EVAL, Pain, NEGAFF, POSAFF, Eval. We applied a relief feature selection measure in order to sort the features according to their a priori usefulness for the items to classify. The results of the feature relief estimation, in descending order, are presented in Table 6.4.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Relief measure</th>
<th>Feature</th>
<th>Relief measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>0.3388</td>
<td>NUMBIGPOSSUM</td>
<td>0.0024</td>
</tr>
<tr>
<td>PATH</td>
<td>0.2311</td>
<td>Hostile</td>
<td>0.0024</td>
</tr>
<tr>
<td>LEX</td>
<td>0.1594</td>
<td>Pleasure</td>
<td>0.0012</td>
</tr>
<tr>
<td>Neg</td>
<td>0.0232</td>
<td>Vice</td>
<td>0.0009</td>
</tr>
<tr>
<td>Virtue</td>
<td>0.0229</td>
<td>EVAL</td>
<td>0.0004</td>
</tr>
<tr>
<td>NUMALL</td>
<td>0.0127</td>
<td>Pain</td>
<td>0.0003</td>
</tr>
<tr>
<td>NUMNEGSUM</td>
<td>0.0111</td>
<td>NEGAFF</td>
<td>0.0002</td>
</tr>
<tr>
<td>NUMPOSSUM</td>
<td>0.0105</td>
<td>POSAFF</td>
<td>0.0001</td>
</tr>
<tr>
<td>Ngtv</td>
<td>0.0048</td>
<td>Eval</td>
<td>0.00007</td>
</tr>
<tr>
<td>NUMBIGNEGSUM</td>
<td>0.0047</td>
<td>BIGGESTNUMOTHER0</td>
<td>-0.0129</td>
</tr>
<tr>
<td>EMOT</td>
<td>0.0043</td>
<td>BIGGESTVALUE</td>
<td>-0.0156</td>
</tr>
<tr>
<td>Pstv</td>
<td>0.0040</td>
<td>BIGGESTSUM</td>
<td>-0.0177</td>
</tr>
<tr>
<td>Pos</td>
<td>0.0040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 6.7 Results of the Subjective Personal Sense Identification

We performed a classification experiment with 475 word pairs with subjective Personal Sense and a random group of 475 objective word pairs. We used NB as the classification algorithm, as other classifiers (SVM and Decision Tree) performed

---

9 We refer to [http://www.wjh.harvard.edu/~inquirer/](http://www.wjh.harvard.edu/~inquirer/) for more information.

worse in our preliminary experiments, probably because of the large number of features and the sparseness of some features' values. We tested different feature combinations, including the ones based on the relief measure threshold. The results are presented in Table 6.5 using three evaluation measures: the F-measures ([Beitzel, 2006]) for objective and subjective groups separately, and the mean accuracy.

<table>
<thead>
<tr>
<th>Features</th>
<th>F-measure for Objective items</th>
<th>F-measure for Subjective items</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.85</td>
<td>0.85</td>
<td>0.86</td>
</tr>
<tr>
<td>POS+WORD+PATH, relief &gt; 0.01</td>
<td>0.86</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>14 features with relief &gt; 0.001</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>22 features with relief &gt; 0.0</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>No LEX</td>
<td>0.70</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>No POS</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>No PATH</td>
<td>0.85</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>No NUMALL</td>
<td>0.86</td>
<td>0.85</td>
<td>0.86</td>
</tr>
<tr>
<td>No NUMPOSSUM</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>Leaving out any other feature</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>LEX+POS+PATH+SentiWN 11 features</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>LEX+POS+PATH+GenInq 17 features</td>
<td>0.87</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>LEX+POS+PATH+SentiWN with positive relief only, 8 features</td>
<td>0.87</td>
<td>0.88</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Almost all the feature sets yielded a considerable performance in terms of accuracy and F-measure: all of them are higher than 0.84. It is only when we leave out the LEX feature, i.e. the potential indicator word, we get a much lower result, but still a very satisfactory one: 73% accuracy, 0.70 and 0.75 F-measure for objective and subjective word pairs respectively. On one hand, it shows that the semantic features derived from the lexical sources make a useful contribution and allow the prediction of the subjective impact of a lexeme successfully. On the other hand, a considerable difference between the results shows that the most effective way to predict a word’s subjective Personal Sense impact in the test set is to learn it in the training set. Evaluating it in principle using the existing lexical resources is less successful.
We performed the experiments, leaving out one of the features every time, in order to evaluate the impact of every individual feature separately. The word serving as the potential Personal Sense indicator, represented by the LEX feature, has proved to play the most important role in the classification: leaving out this feature has decreased the performance most of all. According to our expectations, the syntactic dependency path is also very important, as it also decreased the result comparing to using all the features with a higher than 99% significance in terms of accuracy and both subjective and objective F-measures. The third feature that introduced significant decrease when leaving it out is the part of speech of the indicator (POS), yielding 94-98% statistically significant difference in the result for the three evaluation measures.

6.8 Analysis of Results
It is apparent that there are three most useful features: the lexical item serving as a potential PS indicator (LEX), its part of speech (POS), and the syntactic dependencies path between the indicator and the target noun (PATH). The classification results give additional evidence for this.

It is important to notice that from these three most important features, the part of speech (POS) plays an exceptional role, as in the classification it played a considerably smaller role than LEX and PATH, but it has been identified as the most important feature in terms of the relief measure. We find this result revealing, knowing that the relief measure is based on the ability of the feature in question to discriminate between the items that are similar to each other but belong to different classes. In the actual classification experiment this is not always the case, and the different classes are distinguished in addition by the features from the lexical resources. However, when such an intuition is used to evaluate the features a priori, the part of speech plays a very important role, because a different part of speech for the same word form can make a big difference in terms of subjectivity, especially when most of the other features stay the same.

Not surprisingly, leaving out the features derived from the lexical resources did not change the result dramatically: first of all, our features obviously contained redundant information bringing considerable noise to the dataset; on the other hand, most of the information was contained in the three most useful features, and for the...
closed dataset with a limited vocabulary the LEX feature represented all the lexical information and made the lexical resources features redundant.

The best results overall were achieved in the following two cases: using the three main features plus the General Inquirer features (which had all positive relief), or the three main features with the SentiWN features that had positive relief. Both of these feature sets, and the set consisting of the 22 positive relief features, performed better than the set containing of all the features, with a statistical significance higher than 99.99%. This shows that, first of all, the features characterized by a positive relief actually contribute positively to the classification experiment. On the other hand, it proves that both the SentiWN and the General Inquirer features increase the performance, but can be substituted by one another and do not significantly affect the performance when used together compared to using any one of them.

It is important to notice that the use of the syntactic path as a feature increased the performance considerably, but the absence of this feature did not completely decrease the performance. This supports our contention discussed earlier, confirming that a word often brings subjectivity to the whole sentence, not only to a specific word denoting an object. We realize that this is reasonable if we consider the following examples: “Bombers kill shoppers”, “Mortar assault leaves at least 18 dead”, or “Goal delight for Sheva”. The Personal Sense indicator words are in italics, and the target words are shown in bold. In each case both of the nouns (or parts of speech acting as nouns) occurring in the sentence acquire a subjective Personal Sense, in these cases with the same polarity. However, the emotion contained in the Personal Sense of the different nouns would be different: ‘fear’ for ‘bombers’ and ‘assault’, ‘sadness’ for ‘shoppers’ and ‘18’. It means that the meaning of the syntactic path feature can be more clearly analyzed in experiments with a more fine-grained emotion classification, not a binary subjectivity-objectivity one.

6.9 Conclusions

There has been a vast amount of work done in subjectivity identification. We develop a new subjectivity annotation scheme based on the notion of Personal Sense. First of all, our theoretical motivation was a consideration underlined in [Scherer, 1999], that intentional object is an inherent property of emotion. Secondly, we have shown by analyzing a news headlines sentiment dataset, that a fine-grained approach
to subjectivity is essential, because it allows for a more detailed emotion, subjectivity and polarity annotation, thus enabling more accurate identification of different emotions expressed towards different intentional objects in text.

Thus we suggested a Personal Sense based annotation scheme and applied it to the dataset in question.

In order to classify the resulting subjective and objective expressions automatically, we added to the widely used subjectivity clues approach a syntactic feature, serving as a regular connection between expressed emotion and its intentional object. In a preliminary experiment we showed that the resulting feature set describes the subjectivity identifier are more accurately than the baseline lexical approach.

Finally, we performed a subjectivity-objectivity classification experiment, using the suggested fine-grained sub-sentence level annotation and a set of lexical and additional syntactic features, evaluating the impact of different features and comparing the effect of features derived from two lexical resources.

The fine-grained linguistic approach we have presented here, based on the concept of Personal Sense, is designed for annotating and analyzing automatically subjectivity, polarity and emotions in text. It yields a high performance in the subjectivity classification of word-pairs in the news titles dataset, establishing a useful background for identifying polarity and emotions based on the same annotation scheme. The features introduced in the Personal Sense approach are appropriate for the classification, mostly making a positive contribution to the result. The relative impact of the features is realistically estimated with the chosen relief measure.

Both lexical resources used to infer subjective categories for identifying subjective Personal Sense, SentiWN and General Inquirer, influence the classification result positively, which proves their appropriateness for the task. However, the impact of the derived categories is not high and cannot substitute the usage of the lexical items themselves, at least for the limited dataset available. More experiments on a larger dataset are necessary to investigate the meaning of the resources more thoroughly. Moreover, in the current experiment the use of both resources yielded considerable noise, resulting in a slightly lower classification result. Equally good results were achieved when using one of the two resources separately, which indicates that they are interchangeable in the current setting.
Although the syntactic path feature was meaningful: its absence decreased the subjectivity classification results, it is obvious that the decrease was not dramatic. This goes in line with the intuition described above and followed by most of the research in subjectivity: that sentiment (in our case – subjectivity) is distributed over all the objects in the sentence and in a large number of cases is not restricted to only one noun. However, this fact requires further investigation in a more detailed manner in polarity and emotion classification experiments in the same fine-grained Personal Sense framework.
Chapter 7: Conclusions

Volumes of personal texts in the world wide web are ever-increasing. Whether it is product reviews, opinionated news written by professionals or personal diaries describing lives of ordinary people, there is a lot of information that we can say about authors by analyzing their texts. Personal Sense – defined in [Leontev, 1978] as a former of our subjective consciousness – provides excellent opportunities and relates private states to individual personality.

By harnessing the Personal Sense in blogs, we set out to infer information about the private states and characteristics of bloggers. In this thesis we applied the notion of Personal Sense to a number of experiments in sentiment analysis.

First of all, we applied distributional techniques used to analyze word-meaning, to discover the Personal Sense of the words ‘movie’ and ‘film’ in movie reviews, and infer positive or negative Personal Sense to aim automatic polarity classification of the movie reviews. We came to the conclusion that the Personal Sense of these words and its positive or negative aspects could be presented individually by different authors. We subsequently used the authorship of the reviews and classify documents by different authors each in a separate machine learning experiment. We obtained promising results, which confirmed our assumption that positive or negative affect or appraisal is expressed in a different way by individuals. In order to apply this consideration to a real-life task, where the author is not usually known, we performed the same experiments on groups of texts classified automatically by authorship, also getting positive results and confirming that the individual analysis of polarity can improve the results.
Our next step concerned the automatic classification of authors by their professional background. We assumed that the Personal Sense of a number of words would put them in different semantic relation, for example, semantic proximity to each other, thus constructing a personalized thesaurus containing word-meanings modified by the Personal Sense for every individual author. Our hypothesis was that the personalized thesauri would be similar between the authors of the same professional background and different for the authors of the different background. We successfully tested our hypothesis, using four techniques of Personal Sense representation, with texts by two classes of authors: of military and accounting professional backgrounds.

We concluded that the Personal Sense is in practice a useful notion in subjective language research, and in our final experiment we present an annotation scheme for polarity and emotion recognition in text based on Personal Sense. Using lexical and syntactic techniques of Personal Sense representation, we perform a subjectivity classification on a dataset of news titles, and get very promising results.

The Personal Sense approach has proved to be useful in some tasks in subjectivity analysis in the blogosphere. Accordingly, we applied the notion of Personal Sense to analysis of subjectivity in writings of bloggers, and thus more accurately inferred information about their private states and personal characteristics.

7.1 Future Work

We have demonstrated that identification of Personal Sense can play an important role in opinion mining. Polarity classification, perspective determination, subjective language identification can gain from the advances in the Personal Sense technique. However, further experiments are necessary to establish the precise role of the suggested technique in subjectivity analysis. First of all, it has been shown to be a suitable technique in fine-grained subjective language identification. Further experiments in this direction should demonstrate the Personal Sense technique as an appropriate way of representing and analyzing fine-grained polarity and emotion language, and a useful framework for accurate automatic classification.

Moreover, further investigations on the role of Personal Sense in perspective determination should follow. The suggested Personal Sense structures can represent professional background information about authors of texts. A further step in this...
direction is necessary to develop an algorithm for comparing the obtained structures and obtaining features that would be particularly useful for background classification and perspective determination of authors.

In conclusion, harnessing Personal Sense provides us with important means to say more about bloggers, and investigation should proceed in order to more thoroughly explore these possibilities.
Bibliography


Intelligence and Intelligent Agent Technology, December 9-12, 2008. Sydney, Australia. pp. 320-324.


List of Publications

