CENTRO INTERNACIONAL DE ESTUDOS
DE DOUTORAMENTO E AVANZADOS
DA USC (CIEDUS)

TESIS DE DOCTORADO

CORPUS-BASED CONSTRUCTION OF SENTIMENT LEXICON TO IDENTIFY EXTREME OPINIONS BY SUPERVISED AND UNSUPERVISED MACHINE LEARNING METHODS

Sattam Mohammad Ata Al-Matarneh

ESCUELA DE DOCTORADO INTERNACIONAL

PROGRAMA DE DOCTORADO EN Investigación en Tecnoloxías da Información

SANTIAGO DE COMPOSTELA

2018

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Corpus-based Construction of Sentiment Lexicon to Identify Extreme Opinions by Supervised and Unsupervised Machine learning Methods

SATTAM MOHAMMAD ATA AL-MATARNEH

Presento a miña tese, seguindo o procedemento axeitado ao Regulamento, e declaro que:

1. A tese abarca os resultados da elaboración do meu traballo.
2. De selo caso, na tese faise referencia ás colaboracións que tivo este traballo.
3. A tese é a versión definitiva presentada para a súa defensa e coincide coa versión enviada en formato electrónico.
4. Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.

En SANTIAGO DE COMPOSTELA, 1 de septiembre de 2018

Asdo. SATTAM MOHAMMAD ATA AL-MATARNEH

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AUTORIZACIÓN DOS DIRECTORES DA TESE: Corpus-based Construction of Sentiment Lexicon to Identify Extreme Opinions by Supervised and Unsupervised Machine Learning Methods.

D. Pablo Gamallo

INFORMAN:

Que a presente tese, correspóndese co traballo realizado por **D. SATTAM MOHAMMAD ATA AL-MATARNEH**, baixo a miña dirección, e autorizo a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como director desta non incorre nas causas de abstención establecidas na Lei 40/2015.

En SANTIAGO DE COMPOSTELA, 1 de septiembre de 2018

Asdo. Pablo Gamallo

AUTORIZACIÓN DOS TITOR DA TESE: Corpus-based Construction of Sentiment Lexicon to Identify Extreme Opinions by Supervised and Unsupervised Machine Learning Methods.

D. David E. Losada Carril

INFORMAN:

Que a presente tese, correspóndese co traballo realizado por **D. SATTAM MOHAMMAD ATA AL-MATARNEH**, baixo a miña dirección, e autorizo a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como director desta non incorre nas causas de abstención establecidas na Lei 40/2015.

En SANTIAGO DE COMPOSTELA, 1 de septiembre de 2018

Asdo. David E. Losada Carril

I dedicate this work

To my first inspiration ... my father.

To my refuge when life hardens ... my mother.

My companion in life ..... Nisreen.

To the light of my life ... Salma and Naser.

To my brothers and sisters: Amal, Maher, Samah, Amani, Israa, and Omar.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

And He taught Adam all the names, then showed them to the angels, saying: Inform Me of the names of these, if ye are truthful. They said: Be glorified! We have no knowledge saving that which Thou hast taught us. Lo! Thou, only Thou, art the Knower, the Wise.

Holy Qura'n

When a man dies, his deeds come to an end except for three things: Sadaqah Jariyah (ceaseless charity); a knowledge which is beneficial, or a virtuous descendant who prays for him (for the deceased).

Prophet Mohammad

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Acknowledgments

At the outset, I would like to express my thanks appreciation and gratitude to my supervisor, **Dr. Pablo Gamallo** for the tremendous efforts he has made to accomplish this work. He was always available for help and advice. His valuable comments, remarks, and meetings were a key to all my achievements until now.

He believed in my abilities throughout my PhD journey, and was always supportive and encouraging which helped me reach what I am now.

Special thank also to my tutor Dr. David Losada for his efforts from the beginning of my PhD march.

I am grateful to Dr. Afonso Xavier for his help in the linguistic audit where his guidance and advice greatly influenced the completion of this thesis. Thank you for taking the trouble to help me. I do appreciate it.

I would like to thank all my colleagues at the Centro de Investigación en Tecnoloxías da Información (CiTIUS) at the Universidade de Santiago de Compostela specially Firas Awaysheh. I also would like to thank administration staff in CiTIUS for being supportive whenever I needed them.

I am very grateful to Erasmus Mundus (PEACE II project) that offered me one year grant to complete my PhD within the project PEACE II, which helped me in part of the expenses of my studies. I would like to extend my special thanks to the coordinator of the PEACE II Project, Dr. Sami Ashour, for his efforts.

Words are powerless to express my gratitude to my beloved parents, who always believed in me and supported me with their reassuring words and many prayers. They have always been the main reason for my success and the real spark behind my motivation. I do not exaggerate when I say I would not have achieved this without them.

I also thank my brothers and sisters, Amal, Mahir, Samah, Amani, Israa and Omar for being my friends and soulmates forever.

I want to thank my parents-in-law. You have been a huge support for me, my wife and our children at all times. Special thanks go to my brother-in-law Ashraf Aldmour Who has always been with us in every step, provides the support and assistance despite the distance between us.

I will not forget to thank my all dear friends who have had a great impact, to start this phase of my life. They believed in me and supported me at all times. Particularly, I would like to mention Ala'a Alsarayrah, Yousef Adamantly, Hamzah Almatarneh, Hatem Alameryeen, Ahmad Jawdat, Qasem Obiedat, Shadi Banitaan and Mohammad Al-Abed. Thank you for your continuous support.

Finally, I want to thank my lovely wife, Nisreen for your patience, your strength, and your endless love for our gorgeous family. Your efforts to take care of me, the kids, the house, and your study every day are incredible. Your love for me and our children is limitless, and your permanent support is the reason for my success. I am very lucky to have you in my life. I love you. Thank you!

2018

Resumo

Os estudos en análise de sentimentos e minería de opinións véñense centrando en moitos aspectos relacionados coas opinións, en particular na clasificación de polaridade usando valores positivos, negativos ou neutros. Con todo, a maioría dos estudos pasaron por alto a identificación de opinións extremas (opinións moi negativas e moi positivas), a pesar da súa gran importancia en moitas aplicacións. Esta tese de doutoramento describe unha estratexia para construír léxicos de sentimentos do corpus. Esta estratexia foi utilizada para construír algúns léxicos para coñecer a súa efectividade na determinación da polaridade das opinións. En primeiro lugar, construiremos un léxico específico para cada dominio a partir dun corpus de reseñas de filmes. As palabras de polaridade no léxico están asignadas a pesos que representan diferentes graos de positividade e negatividade. Este léxico combinarase nun sistema de análise de sentimentos para avaliar o seu desempeño na tarefa de clasificación dos sentimentos.

En segundo lugar, dous léxicos serán construídos con palabras moi negativas e positivas a partir de corpus etiquetados. Integraremos os léxicos que se incorporaron aos clasificadores, xa sexan supervisados ou non supervisados. Usaremos un clasificador supervisado, máis precisamente, Support Vector Machine (SVM) con algunhas características lingüísticas como un saco de palabras, incrustación de palabras, léxicos de polaridade e un conxunto de características textuais, co obxectivo de identificar opinións extremas e proporcionar unha análise completa da importancia relativa de cada conxunto de características. Compararemos tamén os nosos léxicos con catro léxicos de sentimentos coñecidos. Para este efecto, realízase unha avaliación indirecta. Os léxicos serán integrados en clasificadores de sentiment supervisados, e o seu rendemento será avaliado en dúas tarefas de clasificación de sentimentos para identificar: a) as opinións máis negativas vs. as que non son moi negativas, e ii) as máis positivas vs. as que non son moi positivas. Ademais, un conxunto de características textuais será integrado nos clasificadores para analizar como estas características textuais melloran o rendemento do léxico. Por outra banda, tamén probaremos a eficacia dos nosos léxicos para determinar opinións extremas utilizando clasificadores sen supervisión. O noso algoritmo de clasificación baséase nun esquema fundamental de combinación de palabras para realizar análises de sentimentos non supervisados.

Palabras chave: Análise de sentimentos, Minería de opinión, Léxico de sentimento, Opinións extremas, Clasificación de polaridade, Aprendizaxe automática

Resumen

Los estudios en análisis de sentimientos y minería de opiniones se vienen centrando en muchos aspectos relacionados con las opiniones, en particular en la clasificación de la polaridad mediante el uso de valores positivos, negativos o neutros. Sin embargo, la mayoría de los estudios han pasado por alto la identificación de opiniones extremas (opiniones muy negativas y muy positivas) a pesar de su gran importancia en muchas aplicaciones. Esta tesis doctoral describe una estrategia para construir léxicos de sentimientos a partir de corpus, en particular en aquellos léxicos con valores extremos. Esta estrategia ha sido utilizada para construir algunos léxicos y para conocer su efectividad en la determinación de la polaridad de las opiniones. Primero, construiremos un léxico específico para cada dominio a partir de un corpus de reseñas de películas. A las palabras de polaridad del léxico se les asignan pesos que representan diferentes grados de positividad y negatividad. Este léxico se combinará en un sistema de análisis de sentimientos para evaluar su desempeño en la tarea de clasificación de sentimientos.

Segundo, se construirán dos léxicos con palabras muy negativas y positivas de corpus etiquetados. Integraremos los léxicos que se han incorporado en los clasificadores, ya sean supervisados o no supervisados. Usaremos un clasificador supervisado, concretamente Support Vector Machine (SVM), con algunas características lingüísticas como una bolsa de palabras, incrustación de palabras, léxicos de polaridad y un conjunto de características textuales, con el fin de identificar opiniones extremas y proporcionar un análisis completo de la importancia relativa de cada conjunto de características. También compararemos nuestros léxicos con cuatro léxicos de sentimientos muy conocidos. Para este propósito, se lleva a cabo una evaluación indirecta. Los léxicos se integrarán en los clasificadores de sentimientos supervisados, y su desempeño se evaluará en dos tareas de clasificación de sentimientos para identificar: i) las opiniones más negativas vs. las que no son muy negativas, y ii) las más positivas vs. las que no son muy positivas. Además, un conjunto de características textuales será integrado en los clasificadores para analizar cómo estas características textuales mejoran el rendimiento del léxico. Por otro lado, también probaremos la eficacia de nuestros léxicos para determinar opiniones extremas mediante el uso de clasificadores no supervisados. Nuestro algoritmo de clasificación se basa en un esquema fundamental de coincidencia de palabras para llevar a cabo análisis de sentimientos sin supervisión.

Palabras clave: Análisis de sentimiento, Minería de opinión, Léxico de sentimiento, Opiniones extremas, Clasificación de polaridad, Aprendizaje automático

Summary

Studies in sentiment analysis and opinion mining focused on many aspects related to opinions, particularly polarity classification by making use of positive, negative or neutral values. However, most studies overlooked the identification of extreme opinions (very negative and very positive opinions) in spite of their vast significance in many applications. This doctoral thesis describes a strategy to build sentiment lexicons from corpora, namely lexicons lexicons adapted to extreme values. This strategy has been used to build some lexicons and to know its effectiveness in determining the polarity of opinions. First, we will construct a domain-specific lexicon from a corpus of movie reviews. Polarity words of the lexicon are assigned weights standing for different degrees of positiveness and negativeness. This lexicon is will be combined into a sentiment analysis system to evaluate its performance in the task of sentiment classification.

Second, two lexicons will be built of extremely negative and positive words from labeled corpora. We will integrate the lexicons that have been built into classifiers, whether supervised or unsupervised classifier. We will use a supervised classifier, more precisely, Support Vector Machine (SVM) with some linguistic features such as a bag of words, word embedding, polarity lexicons, and set of textual features, in order to identify extreme opinions and provide a comprehensive analysis of the relative importance of each set of features. We also will compare our lexicons with four well-known sentiment lexicons. For this purpose, an indirect evaluation is carried out. The lexicons will be integrated into supervised sentiment classifiers, and their performance is evaluated in two sentiment classification tasks to identify i) the most negative vs. not most negative opinions, and ii) the most positive vs. not most positive. Moreover, a set of textual features is integrated into the classifiers to analyze how these textual features improve the lexicon performance. On the other hand, we also tested the efficiency of our lexicons in determining extreme opinions through the use of unsupervised classifiers. Our classification algorithm is based on a fundamental word-matching scheme to carry out unsupervised sentiment analysis.

Keywords: Sentiment Analysis, Opinion Mining, Sentiment Lexicon, Extreme Opinions, Polarity Classification, Machine Learning.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Resumen extendido

La revolución de la información es la característica más destacada de este siglo. El mundo se ha convertido en una pequeña aldea con la proliferación de redes sociales donde cualquier persona en todo el planeta puede vender, comprar o expresar sus opiniones. La gran cantidad de información en Internet se ha convertido en una fuente de interés para numerosos trabajos, ya que ofrece una excelente oportunidad para extraer información y organizarla según las necesidades particulares.

A partir del uso masivo de Internet y de las redes sociales en varios aspectos de la vida, estos han llegado a desempeñar un papel importante en la orientación de las tendencias de la gente en los ámbitos social, político, religioso y económico, a través de las opiniones expresadas por los individuos.

Las redes sociales y sus herramientas (por ejemplo, Tweeter, Facebook, LikedIn, etc.) proporcionan información sobre cómo se siente la gente acerca de las cosas que están a su disposición. Además, las organizaciones han acumulado una cantidad significativa de datos sobre cómo piensan sus empleados o clientes en relación a los productos y servicios que reciben. Incluso los departamentos de Recursos Humanos están interesados en hacer seguimiento de la lealtad de los posibles empleados, ya sea para que se conviertan en miembros permanentes de la empresa o para que se vayan después de recibir la consecuente indemnización.

En la última década, se ha publicado un número considerable de estudios en el campo de la minería de opiniones y el análisis de sentimientos. La motivación detrás de estos estudios fue el intento de extraer información útil para ser usada en muchos dominios a partir de la gran cantidad de opiniones disponibles de los usuarios en blogs, redes sociales, noticias y sitios web de compras.

A la vanguardia de todos los demás campos, la Inteligencia de Negocios es el dominio más atractivo para la minería de opiniones, con muchos estudios concentrados en la minería

de comentarios de clientes para un mejor entendimiento del mercado. Otro campo tradicional es el de la inteligencia gubernamental, que se centra en cuestiones como las elecciones, la reputación de los partidos y la elección de políticas de acuerdo con las opiniones de la gente.

Según Pang et al. (2008), el 73% y el 87% de los lectores de reseñas online (restaurantes, hoteles, agencias de viajes o médicos), afirman que las reseñas tuvieron una influencia significativa en su compra.

El Análisis de Sentimientos también llamado Opinion Mining se define como el campo de estudio que analiza las opiniones, sentimientos, evaluaciones, actitudes y emociones de las personas a partir del lenguaje escrito. Es una de las áreas de investigación más activas en el Procesamiento del Lenguaje Natural (PNL) y también es ampliamente estudiada en minería de datos, minería Web y minería de texto (Liu, 2012).

Debido a su importancia para la política, los negocios y la sociedad en su conjunto, la investigación de Análisis del Sentimiento se ha expandido fuera de la informática a otras ciencias como la política, las ciencias sociales y la gestión. La creciente importancia del Análisis de Sentimientos está correlacionada con el crecimiento de las reseñas en línea, foros de discusión, blogs, micro-blogs, Twitter y redes sociales. Por primera vez en la historia de la humanidad, ahora tenemos un volumen masivo de datos de opinión registrados en forma digital para su análisis (Liu, 2012).

La tarea fundamental en Opinion Mining es la clasificación de la polaridad (Pang and Lee, 2008; Cambria, 2016; Cambria et al., 2013), que ocurre cuando un texto que declara una opinión se clasifica en un conjunto predefinido de categorías de polaridad (por ejemplo, positivo, neutro, negativo). Reseñas con "pulgares hacia arriba" versus "pulgares hacia abajo", o "me gusta" versus "no me gusta" son ejemplos de la clasificación de polaridad en dos clases. Una manera inusual de realizar el análisis de sentimientos es detectar y clasificar las opiniones que representan las opiniones más negativas y más positivas sobre un tema, un objeto o un individuo. Las llamamos *opiniones extremas*.

Una opinión extrema es la peor o la mejor visión, juicio o valoración que se forma en la mente de una persona sobre un asunto en particular.

Las opiniones extremas son el foco de atención de las organizaciones o individuos más que otras opiniones estándar. En el caso de cualquier producto, la gente siempre quiere saber los aspectos más negativos para poder evitarlos o solucionarlo. Al mismo tiempo, los clientes siempre quieren comprar el mejor producto, por lo que intentan encontrar aquellos clasificados con 5 estrellas.

Una de las principales motivaciones para detectar opiniones extremas es el hecho de que en realidad representan opiniones *realmente* positivas o negativas. Son opiniones *puras*, sin amigüedad. Como los sistemas de valoración no tienen límites claros dentro de una escala continua, las opiniones débilmente polarizadas (por ejemplo, las calificadas como 4 y 2 en un sistema de valorización de 1 a 5) pueden estar de hecho más cerca de las afirmaciones neutrales. Según Pang and Lee (2005), "es bastante difícil calibrar correctamente las escalas de diferentes autores ya que el mismo número de estrellas, incluso dentro de lo que es ostensiblemente el mismo sistema de puntuación, puede significar cosas diferentes para diferentes autores". Dado que los sistemas de valoración se definen en una escala subjetiva, solo las opiniones extremas pueden ser vistas como afirmaciones positivas/negativas naturales, transparentes y no ambiguas. Las opiniones extremas solo constituyen una pequeña parte de las opiniones sobre los medios sociales. Según Pang and Lee (2005), apenas alrededor del 5% de todas las opiniones se encuentran en los puntos más extremos de una escala, lo que hace que la búsqueda de estas opiniones sea una tarea complicada y muy desafiante.

La literatura sobre Opinion Mining and Sentiment Analysis ha ignorado en su mayoría las opiniones extremas a pesar de su importancia cuando el objetivo es identificar las debilidades y fortalezas más relevantes de cada producto u organización desde el punto de vista de los clientes. Los puntos de vista más negativos ayudan a determinar los aspectos más molestos de los productos para los clientes y cuáles son los productos defectuosos. Por otro lado, los puntos de vista muy positivos permiten la identificación y selección de productos, servicios y vendedores destacados. Además, las opiniones pueden ser indicativas del fraude que practican algunas organizaciones, en particular cuando escriben reseñas muy positivas sobre sí mismas para elevar su calificación. Del mismo modo, estas reseñas y comentarios en redes sociales también se utilizan para desacreditar un producto o servicio, ya que algunos competidores pueden escribir revisiones muy negativas para reducir las ventas de sus competidores haciendo así una especie de competencia desleal, como se menciona en Luca and Zervas (2016).

No es sorprendente que las opiniones tengan un fuerte impacto en las ventas de productos, ya que influyen en las decisiones de los clientes antes de comprar. Estudios anteriores analizaron esta relación, para mostrar que a medida que aumenta la alta proporción de comentarios negativos de los consumidores en línea, también aumenta la actitud negativa del consumidor (Lee et al., 2008). Se han observado efectos similares en las reseñas de los consumidores: las reseñas de una estrella perjudican significativamente las ventas de libros en Amazon.com Chevalier and Mayzlin (2006). El impacto de las revisiones de una estrella,

que representan las opiniones más negativas, es mayor que el impacto de las revisiones de cinco estrellas en este sector del mercado en particular. Los consumidores informan que están dispuestos a pagar entre un 20% y un 99% más por un artículo o servicio que tenga una calificación de 5 estrellas, lo que significa que les gusta pagar hasta el doble de precio por productos con una calificación de 5 estrellas en comparación con el producto con calificación de 4 estrellas (Pang et al., 2008; comScore and Kelsey, 2007; Horrigan, 2008).

Por último, pero no por ello menos importante, otra motivación para la identificación de opiniones es el uso actual de la tecnología bot por parte de los cyborgs en las redes sociales. Estos bots están diseñados para vender productos o atraer clics, amplificando historias falsas o sesgadas con el fin de influir en la opinión pública.

Existen dos enfoques principales para encontrar la polaridad de sentimientos a nivel de documento o de frase. En primer lugar, técnicas de aprendizaje automático basadas en corpus de entrenamiento anotados con información de polaridad y, en segundo lugar, estrategias basadas en léxicos de polaridad. Los enfoques basados en el léxico son muy populares en el análisis de sentimientos y la minería de opiniones, y desempeñan un papel clave en todas las aplicaciones en este campo. La principal preocupación de los enfoques basados en el léxico es que la mayoría de las palabras de polaridad dependen del dominio, ya que el estado de subjetividad de la mayoría de las palabras es muy ambiguo. La misma palabra puede tener una carga subjetiva en un dominio específico, mientras que puede referirse a información objetiva en otro dominio. De ello se deduce que los léxicos dependientes del dominio deberían superar a los diccionarios de propósito general en la tarea de análisis de sentimientos. Sin embargo, la construcción de léxicos de polaridad dependiente del dominio es una tarea ardua y aburrida si se realiza manualmente para cada dominio de destino. Con el crecimiento de los corpus de sentimientos en diversas áreas, la generación automática de este tipo de recursos se está convirtiendo en una tarea fundamental en la minería de opiniones y en el análisis de sentimientos (Huang et al., 2014).

Los objetivos principales de esta tesis son los siguientes:

– Proponer un método para construir automáticamente léxicos de polaridad a partir de corpus. Más específicamente, este método propuesto debe ser capaz de construir léxicos que puedan adaptarse a todos los dominios de datos y que sean aplicables a todas las tareas de clasificación de polaridad.

– Utilizar el método automático basado en el corpus para construir léxicos de opinión de dominio específicos e investigar la efectividad de estos léxicos comparándolos con otros léxicos existentes.

– Utilizar el método basado en corpus para construir un léxico de opiniones extremas, distinguiendo los términos más negativos y más positivos de las otras palabras de opinión.

– Examinar la eficacia de la construcción automática de un léxico de sentimientos mediante un procedimiento de evaluación indirecta. La evaluación indirecta consiste en medir el rendimiento de los clasificadores de aprendizaje supervisado de máquinas basados en el léxico.

– Examinar la eficacia y las limitaciones de diferentes características (*features*) lingüísticas para identificar opiniones mediante el uso de un método de aprendizaje supervisado.

– Investigar la efectividad de la construcción automática de un léxico de sentimientos utilizando la clasificación de aprendizaje automático no supervisado para buscar opiniones.

Estas tareas deben realizarse siguiendo estos pasos:

– La primera fase consiste en proponer un método automático basado en corpus para construir un léxico de opiniones extremas, distinguiendo los términos más negativos y más positivos de las otras palabras de opinión.

– En la segunda fase, identificamos los métodos de aprendizaje supervisado existentes (por ejemplo, máquinas de vectores de apoyo (SVM)) que se pueden aplicar para clasificar datos a escala. Dentro de esta etapa, también identificaremos las características lingüísticas (por ejemplo, N-grams Features, Word Embedding, Sentiment words, etc.), que son una buena guía para extraer comentarios extremos.

– La tercera fase consiste en implementar algoritmos de búsqueda y aprendizaje utilizando las plataformas y herramientas disponibles (por ejemplo, liblinear, libsvm, scikit-learn, etc.).

– En la cuarta fase, evaluaremos la eficacia de los léxicos que hemos desarrollado. Esto se hará comparando nuestros léxicos con otros léxicos bien conocidos (por ejemplo, SO-CAL, SentiWords, AFINN-111, etc.) sobre puntos de referencia estándar que han

sido construidos a lo largo de los años para facilitar la investigación experimental en Minería de Opinión y Análisis de Sentimientos (por ejemplo, Multi-Domain Sentiment Dataset, Large Movie Review Dataset, etc.).

– Para medir la eficiencia y la eficacia de los modelos y léxicos, en la quinta fase utilizaremos la precisión y el recall, que son dos medidas comunes, para evaluar la eficacia de outputs recurrentes. Precisión (P), es la fracción de objetos recuperados que son relevantes. Recall (R), es la fracción de objetos relevantes que son recuperados por el sistema, también utilizaremos la medida F, que es una medida única que compensa la precisión con el recall, que es la media armónica ponderada de precisión y recall. Además, para determinar el rendimiento de los algoritmos, utilizaremos la significación estadística.

– En la sexta fase, repetiremos todas las fases anteriores para el aprendizaje no supervisado.

Contents

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 1

Introduction

The information revolution is the most prominent feature of this century. The world has become a small village with the proliferation of social networking sites where anyone around the planet can sell, buy or express their opinions. The vast amount of information on the Internet has become a source of interest for studies, as it offers an excellent opportunity to extract information and organize it according to the particular needs.

After the massive explosion in the use of the Internet and social media in various aspects of life, social media has come to play a significant role in guiding people's tendencies in social, political, religious and economic domains, through the opinions expressed by individuals.

The social media and its tools (i.e., Tweeter, Facebook, LikedIn, etc.) provide information on how people feel about things available to them. Also, organizations have accumulated a significant amount of data about how their employees or customers think about the products and services they receive. Even Human Resources departments are interested in understanding the loyalty of prospective employee whether they will become long-term members of the company or would leave after receiving training and benefits.

In the last decade, a considerable number of studies has been published in the field of opinion mining and sentiment analysis. The motivation behind these studies was the attempt to extract useful information to be used in many domains from the vast amount of available users views on blogs, social networks, news, and shopping websites.

At the forefront of all other fields, Business Intelligence is the most attractive domain for opinion mining with many studies concentrated on mining customers reviews for better market understanding. Another traditional field is government intelligence, which focuses

on issues such as elections, parties reputation, and choosing policies according to people opinions.

According to Pang et al. (2008), 73% and 87% among readers of online reviews such as (restaurants, hotels, travel agencies or doctors), state that reviews had a significant influence on their purchase.

Sentiment Analysis also called Opinion Mining is defined as the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in Natural Language Processing (NLP) and is also widely studied in data mining, Web mining, and text mining (Liu, 2012).

Because of its importance to political, business and society as a whole, Sentiment Analysis research has expanded outside computer science to other sciences such as management, political and social studies. The increasing importance of Sentiment Analysis is correlated with the growth of online reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. For the first time in human history, we now have a massive volume of opinionated data recorded in digital form for analysis (Liu, 2012).

The fundamental task in Opinion Mining is polarity classification (Pang and Lee, 2008; Cambria, 2016; Cambria et al., 2013), which occurs when a piece of text stating an opinion is classified into a predefined set of polarity categories (e.g., positive, neutral, negative). Reviews such as "thumbs up" versus "thumbs down", or "like" versus "dislike" are examples of two-class polarity classification. An unusual way of performing sentiment analysis is to detect and classify opinions that represent the most negative and most positive opinions about a topic, an object or an individual. We call them *extreme opinions*.

An extreme opinion is the worst or the best view, judgment, or appraisal formed in one's mind about a particular matter.

Extreme opinions are the focus of attention for organizations or individuals more than other standard opinions. About commodities with low ratings, people always want to know the worst aspects about goods, services, places, etc. so that they can avoid them or fix them. At the same time, as customers always want to buy the best product, they try to find the 5-stars rated products.

1.1 Motivation

One of the main motivations for detecting opinions is the fact that they actually stand for *pure* positive and negative opinions. As rating systems have no clear borderlines on a continuum scale, weakly polarized opinions (e.g. those rated as 4 and 2 in a 1 to 5 rating system) may be, in fact, closer to neutral statements. According to Pang and Lee (2005), "it is quite difficult to properly calibrate different authors' scales, since the same number of *stars* even within what is ostensibly the same rating system can mean different things for different authors". Given that rating systems are defined on a subjective scale, only opinions can be seen as natural, transparent, and non ambiguous positive / negative statements. Fig. 1.1 shows the expected distribution of negative, neutral and positive opinions on a scale from 1 to 5. Red, blue, and green colors stand for negative, neutral and positive opinions, respectively. Color overlap covers the space around 2 and 4, where neutral views may appear together with light negative and positive opinions. Pure red and green appear only around 1 and 5 stars, representing the extreme opinions.

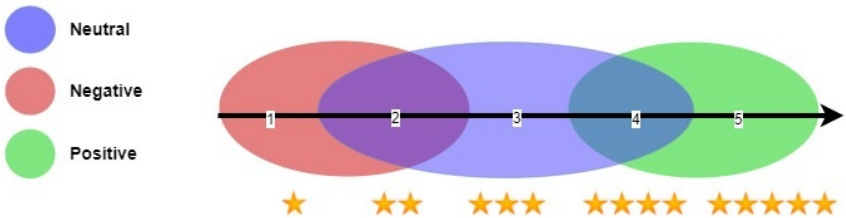

Figure 1.1: Hypothetical continuous distribution of negative, neutral and positive views on a scale from 1 to 5, according to the borderline between stars.

Extreme opinions only constitute a small portion of the opinions on Social Media. According to Pang and Lee (2005), only about 5% of all opinions are on the most extreme points of a scale, which makes the search for these opinions a very challenging task.

The literature on Opinion Mining and Sentiment Analysis has mostly ignored extreme opinions in spite of their importance when the objective is to identify the most relevant weaknesses and strengths of each product or organization from the viewpoint of customers. The most negative viewpoints help to determine the most annoying aspects of products for cus-

tomers and what are the defective goods. On the other hand, strongly positive views allow for the identification and selection of outstanding products, services, and sellers. Also, views may be indicative of fraud practised by some organizations, namely when they write very positive online reviews about themselves to raise their rating. Similarly, these reviews are also used to discredit a product or service, since some competitors may write very negative reviews to reduce the sales of their competitors as a kind of unfair competition, as mentioned in Luca and Zervas (2016).

It is not surprising that views have a strong impact on product sales since they influence customer decisions before buying. Previous studies analyzed this relationship, to show that as the high proportion of negative online consumer reviews increases, so does the consumer's negative attitude (Lee et al., 2008). Similar effects have been observed in consumer reviews: one-star reviews significantly hurt book sales on Amazon.com (Chevalier and Mayzlin, 2006). The impact of 1-star reviews, which represent the most negative views, is greater than the impact of 5-star reviews in this particular market sector. Consumers report that they would be willing to pay from 20% to 99% more, depending on the type of item or service that rated 5-star than the same one rated as 4-star, which means that they like to pay up to double price more on products with a five-star rating compared to the four-star rating product (Pang et al., 2008; comScore and Kelsey, 2007; Horrigan, 2008).

Last but not least, another motivation for the identification of extreme opinions is the current use of bot technology by cyborgs on social networks. These bots are designed to sell products or attract clicks, amplifying false or biased stories in order to influence public opinion.

As for the motivations that are behind the need to automatically build polarity lexicons, there exist two main approaches to finding the sentiment polarity at document or sentence level. First, machine learning techniques based on training corpora annotated with polarity information and, second, strategies based on polarity lexicons. Lexicon-based approaches are very popular in sentiment analysis and opinion mining, and they play a key role in all applications in this field. The main concern of lexicon-based approaches is that most polarity words are domain-dependent since the subjectivity status of most words is very ambiguous. The same word may be provided with a subjective burden in a specific domain while it can refer to objective information in another domain. It follows that domain dependent lexicons should outperform general-purpose dictionaries in the task of sentiment analysis. However, the construction of domain-dependent polarity lexicons is a strenuous and boring task if it is

made manually for each target domain. With the growth of sentiment corpora in diverse areas, the automatic generation of this kind of resources is becoming a fundamental task in opinion mining and sentiment analysis (Huang et al., 2014).

1.2 Problem Statement

There is a need for systematic studies attempting to understand how to mine the vast amount of unstructured text data in order to extract comments and opinions. Previous studies have considered that, in whatever rating system, it is possible to identify three categories: negative, neutral, and positive views. For instance, on a 5-rating scale, negative opinions are those that belong to the reviews of one and two stars, positive views are those assigned four and five stars, while three-star is neutral.

Since the dictionaries play a crucial role in polarity classification task, literature offered many lexicons that have been constructed in different ways, but none of these lexicons has been devoted to words denoting extreme sentiments.

By contrast, our thesis relies on two binary classification tasks focused on identifying opinions. First, our objective is to build a classifier identifying the most negative views against other opinions, including not very negative, neutral, and positive. Secondly, we also define a classifier, selecting the most positive views from the rest of opinions, namely those that are not very positive, neutral, and negative. The key aspect of our strategy is based on the construction of the polarity lexicon underlying this type of classification.

1.3 Objectives

The main objectives of this thesis are the following:

- To propose a method for automatically building polarity lexicons from corpora. More specifically, this proposed method must be capable of building lexicons that can be adapted to all domains as well as being applicable to all polarity classification tasks.

- To use the automatic corpus-based method to build specific domain opinion lexicons and investigate the effectiveness of this lexicon by comparing it with other popular lexicons.

– To use the corpus-based method to build a lexicon of extreme opinions, by distinguishing the most negative and most positive terms from the other opinion words.

– To examine the effectiveness of the automatic construction of a sentiment lexicon using an indirect evaluation procedure. The indirect evaluation consists of measuring the performance of supervised machine learning classifiers based on the lexicon.

– To examine the effectiveness and limitations of different linguistic features to identify opinions by using a supervised learning method.

– To investigate the effectiveness of the automatic construction of a sentiment lexicon using unsupervised machine learning classification to search for opinions.

1.4 Methodology

The main objective of our study is to search for extreme opinions by a corpus-based construction of sentiment lexicon with different machine learning algorithms (supervised and unsupervised) and textual features and among varying levels of training data.

This task needs to be done by following these steps:

– the first phase is to propose an automatic corpus-based method to build a lexicon of extreme opinions, by distinguishing the most negative and most positive terms from the other opinion words.

– The second phase aims at identifying existing supervised learning methods (e.g., Support Vector Machines (SVM)) that can be applied to classify scaled data. Within this stage, we will also identify the linguistic features (e.g., N-grams, Word Embedding, Sentiment words, etc.), that are a good guidance for extracting extreme comments.

– The objective of the third phase is to implement search and learning algorithms using available platforms and tools (e.g., liblinear, libsvm,scikit-learn, etc.).

– In the fourth phase, we evaluate the efficiency of the lexicons that we have developed. This will be done by comparing our lexicons with other well-known lexicons (e.g., SO-CAL, SentiWords, AFINN-111, etc.) on standard benchmarks that have been constructed over the years to facilitate experimental research in Opinion Mining and Sentiment Analysis (eg, Multi-Domain Sentiment Dataset, Large Movie Review Dataset, etc).

– In order to measure the efficiency and effectiveness of the models and lexicons, in the fifth phrase, we use precision and recall, which represent two common measures for assessing the effectiveness of recurring outputs. Precision is the fraction of retrieved objects that are relevant. Recall is the fraction of relevant objects that are retrieved by the system. Also, we use F-measure, which is the weighted harmonic mean of precision and recall. Moreover, to determine the performance of algorithms, we make use of statistical significance. Finally, we analyze the results obtained in the fourth phase.

– In the sixth phase, we will repeat all previous phases for unsupervised learning.

1.5 Outlines

This thesis consists of six other chapters. In Chapter 2 we introduce a background of the related literature. We review the main tasks and levels in Sentiment Analysis and Opinion Mining. Also, we discuss the main approaches and techniques in sentiment classification by study the historical context and current practice. We also offer evaluation methods and the set resources that were used in the current study, such as datasets and lexicons.

Our proposed method of building sentiment lexicons was described in Chapter 3. The most significant linguistic features and its representation that were employed in our experiments have been demonstrated in Chapter 4.

In chapter5, we reported an extensive set of experiments aimed to compare our automatic construction lexicon with other four well-known handcraft lexicons for three binary classification tasks:

– Positive vs. negative

– very negative(VN) vs. not very negative opinions (NVN),

– very positive (VP) vs. not very positive opinions (NVP).

Also, we examined the effectiveness and limitations of different linguistic features to identify extreme opinions.

In Chapter 6, we investigated the effectiveness of the automatic construction of a sentiment lexicon using unsupervised machine learning classification to search for extreme opinions.

Last but not least, Chapter 7 includes the conclusions of this thesis and recommends future lines of work.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CHAPTER 2

BACKGROUND

Due to the huge number of papers devoted to Sentiment Analysis, several surveys and books have been published to review all topics and tasks in the field. Pang et al. (2008) conducted a general survey of more than three hundred papers by reporting applications, common challenges for sentiment analysis. Liu and Zhang (2012) and Liu (2015) performed a survey of different tasks and works published in Sentiment Analysis and Opinion Mining. Tsytsarau and Palpanas (2012) presented a survey on Sentiment Analysis by focusing on opinion mining, opinion aggregation including spam detection and contradiction analysis. They compared opinion mining methods that were employed on some common dataset. Cambria et al. (2013) pointed out complexities derived from Sentiment Analysis concerning current strategies along with possible future research directions. Ravi and Ravi (2015) presented a comprehensive, state-of-the-art review of the research work done in various aspects of Sentiment Analysis during 2002–2014. More recently, Hemmatian and Sohrabi (2017) reviewed the opinion mining area and its related classification techniques and developed a survey by examining the well-known existing methods and their challenges. Other relevant surveys and books are those of (Kaur and Gupta, 2013; Silva et al., 2016; Medhat et al., 2014; Soleymani et al., 2017; Liu and Zhang, 2012; Feldman, 2013; Serrano-Guerrero et al., 2015)

In this chapter, we will present a review of the literature on Sentiment Analysis. How to build sentiment lexicons and related issues will be discussed in sections 2.3.2 and 2.4. On the other hand, classification methods of Sentiment Analysis based on machine learning algorithms will be the focus of Section 2.3. Finally, evaluation methodology along with a brief description of existing lexicons and datasets are reported in 2.5.

2.1 Sentiment Analysis Tasks

Many surveys define Sentiment Analysis and Opinion Mining as consisting of several different tasks. Pang et al. (2008) summed the main tasks of Sentiment Analysis into opinion extraction, sentiment classification, polarity determination, and summarization, while Liu and Zhang (2012) listed subjectivity and sentiment classification, aspect-based sentiment Analysis, sentiment lexicon construction, opinion summarization, analysis of comparative opinions, opinion search and retrieval, opinion spam detection and quality of reviews, as major tasks in Sentiment Analysis. For Ravi and Ravi (2015), seven broad dimensions refer to tasks to be accomplished on Sentiment Analysis, namely, subjectivity classification, sentiment classification, review usefulness measurement, lexicon creation, opinion word and product aspect extraction, opinion spam detection and various applications of opinion mining.

In this section, we list the following major tasks: Subjectivity Detection, Opinion Spam Detection, Opinion Summarization, Opinion Polarity Classification, and Lexica and corpora creation (Fig. 2.1).

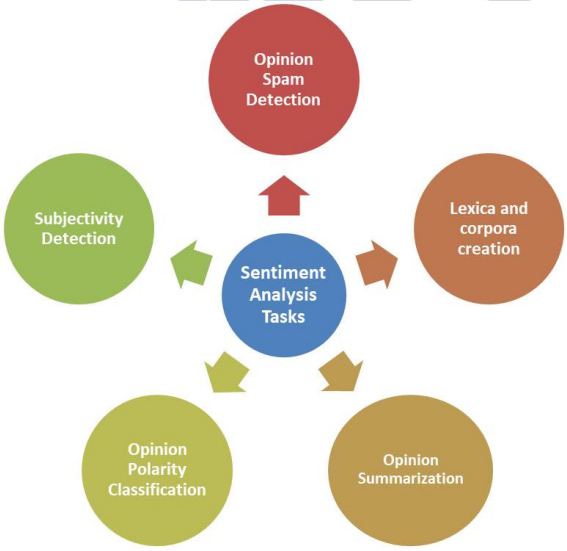

Figure 2.1: Different tasks of Sentiment Analysis.

2.1.1 Subjectivity Detection

Subjectivity Detection is the task of detecting if a text is objective or subjective. Objective texts carry some factual information, for example, "the sky is blue," while subjective texts express somebody's personal views or opinions, for example, "I like the color blue" (Liu and Zhang, 2012). The task of determining whether a sentence is subjective or objective is called subjectivity classification. Wiebe and Riloff (2005) introduced the results of developing subjectivity classifiers using only unannotated texts for training. They provided a list of subjectivity cues with over 8,000 entries. Tang et al. (2009) discussed four problems associated with opinion mining, one of these is subjectivity classification. They highlighted some approaches like Naive Bayes (NB), Multiple Naive Bayes (MNB), and cut-based classifiers. Wang et al. (2011) performed subjectivity classification by considering improved Fisher's discriminant ratio based feature selection method. Experiments were performed on two Chinese corpora, multi-domain reviews, and car reviews. The proposed feature sets along with words appearing in positive (+ve) and negative (-ve) texts were used for training Support Vector Machine (SVM), which yielded sentiment classification accuracy of 86.6%. Benamara et al. (2011) suggested subjectivity classification at the segment level for discourse-based sentiment analysis. Each segment is classified into four classes, S, OO, O, and SN, where S means subjective and evaluative being positive or negative; OO segments are positive or negative opinion implied in an objective segment; O segments contain neither a lexicalized subjective term nor an implied opinion; SN segments are subjective, but non-evaluative (no positive or negative sentiment). Rustamov et al. (2013) described two different supervised machine learning approaches: a Fuzzy Control System and Adaptive Neuro-Fuzzy Inference System, and applied them to sentence-level subjectivity detection in a movie review. Chenlo and Losada (2014) presented a systematic study of different sentence features for two tasks in sentiment classification: namely, polarity and subjectivity classification. They found that unigrams or bigrams combined with sentiment lexicon features consistently give good performance for subjectivity and polarity classification.

Drawing on these results, our proposed method of constructing sentiment lexicons, which we discuss in Chapter 3, may be a useful evidence in determining subjectivity as well as polarity.

2.1.2 Opinion Spam Detection

Opinion spam detection aims to identify fake reviews and fake reviewers (Liu, 2012). The goal is to detect fake opinions in favor of or against a product or service written intentionally by malicious users to make their target popular or unpopular. Jindal and Liu (2008) is one of the first attempts with promising results in this area of study. Ott et al. (2012) presented a general framework for estimating the prevalence of deception in online review communities, based on the output of a noisy deception classifier. They used this framework to explore the prevalence of deception among positive reviews in six popular online review communities. Also, in the same frame, the study of Mukherjee et al. (2012) adopted frequent pattern mining to find groups of reviewers who frequently write reviews together. They then construct features to find the most likely groups of spammers. Mukherjee et al. (2011) built graph modelling relations between groups of spammers, spammers, and products for group spammer ranking. Further, Wang et al. (2012) proposed a new concept of review graph to capture the relationships between the reviews and their corresponding authors as a heterogeneous graph. Mukherjee et al. (2013) developed a novel and principled method to employ observed reviewing behaviors to detect opinion spammers.

Although this task is not considered one of the objectives of our thesis, it may be helpful in identifying counterfeit reviews because they are often written to reduce or increase the value of goods or services. Extreme reviews are fertile environments for this kind of counterfeiting because of their influence on consumer opinions.

2.1.3 Opinion Summarization

Opinion summarization is the task of summarizing a large group of opinions toward a topic, encompassing different perspectives, aspects, and polarities. This is particularly important when someone wants to make a decision because a single opinion cannot be trustworthy. This task extracts the main features that an entity shared within one or several documents and the sentiments regarding them (Wang et al., 2013). Thus, two perspectives can be distinguished: single-document and multi-document summarization. Single-document summarization consists in analyzing internal facts present in the analyzed document, and mainly showing those pieces of texts which better describe them. On the other hand, in multi-document summarization, once features and entities have been identified, the system has to group or order the different sentences which express sentiments related to those entities or features. The final

summary can be displayed as a graphic or a text showing the main features/entities and quantifying the sentiment around each one in some way. Hu and Liu (2004) returns all the negative and positive sentences for each extracted product feature, and a count is given to show the number of positive and negative opinions for each feature. Meng and Wang (2009) addressed the most repeated terms or phrases as the summary of a product feature. Lu et al. (2009) presented a view of aspect ratings for each product. Nishikawa et al. (2010) proposed a novel algorithm for opinion summarization that takes into account content and coherence, concurrently. They directly search for the optimum sentence sequence by extracting and ordering sentences present in the input document set. Lpez Condori and Salgueiro Pardo (2017) introduced a new content selection strategy to produce extractive summaries. Also, they presented a novel Natural Language Generation (NLG) template-based system to generate abstractive summaries of opinions.

The fact that extreme reviews are the most influential in customer decisions make extreme opinions very valuable and exciting in opinion summarization, notably because this task is mainly aimed to summarize views about an object to facilitate decision-making. Even if it is not their primary goal, our lexicons, which were built relying upon the approach proposed in Chapter 3, combined with the linguistic features described in Chapter 4, may be also useful to summarize extreme opinions.

2.1.4 Opinion Polarity Classification

Opinion polarity classification is the task of determining whether the text expresses positive or negative (or sometimes neutral) opinion. As mentioned above, the fundamental task in Opinion Mining is polarity classification (Liu, 2012), which occurs when a piece of text stating an opinion is classified into a predefined set of polarity categories (e.g., positive, neutral, negative). Reviews such as "thumbs up" versus "thumbs down", or "like" versus "dislike" are examples of two-class polarity classification.

As this task is going to be a centerpiece of this thesis, Section 2.3 discusses in more details many of the previous related works.

2.1.5 Lexica and corpora creation

A lexicon is a vocabulary of sentiment words with both sentiment polarity and strength value. All topics related to the sentiment lexicon and the methods of construction of sentiment lexicons were discussed in detail in Sections 2.3.2 and 2.4.

2.2 Levels of Sentiment Analysis

Sentiment analysis typically works at four different levels of granularity, namely document level, sentence level, aspect level, and concept level, as shown in Fig. 2.2. In this section, we will describe in details each of these levels in the following subsections.

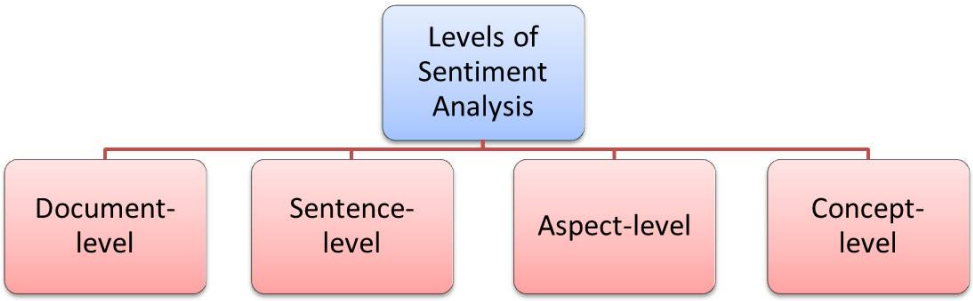

Figure 2.2: Different levels of Sentiment analysis .

2.2.1 Document-level

Document-level is working with whole documents as the basic information unit. It is, hence, the most abstract level of sentiment analysis and thus not appropriate for precise evaluations (Moraes et al., 2013). The result at this level is often general information about the documents. Polarity sentiments are eventually summarized on the whole of the document as positive or negative (Pang et al., 2002). Most early studies in Sentiment Analysis (Turney, 2002; Pang et al., 2002) put their focus at document level and relied on datasets such as movie

and products reviews. After the widespread of the Internet and e-commerce boom, different types of datasets have been collected from websites about customer opinions. The review document often expresses opinions on a single product or service and was written by a single reviewer. These datasets have led to growing number of studies over the years (Seki et al., 2009; Coussement and Van den Poel, 2009; Dang et al., 2010; Bollen et al., 2011; Xia et al., 2011; Moreo et al., 2012; Bosco et al., 2013; Moraes et al., 2013; Basari et al., 2013; Khan et al., 2014; Cho et al., 2014). More than 45 % of the articles published in Sentiment Analysis until 2015 are on the document level (Ravi and Ravi, 2015).

2.2.2 Sentence-level

Since the outcome of Sentiment Analysis at document level is general and does not provide accurate information, and there is a need for deeper analysis, many studies have begun to use the sentences within the document as an approach to analyze opinions (Nasukawa and Yi, 2003; Hiroshi et al., 2004; McDonald et al., 2007; Abbasi et al., 2008; Narayanan et al., 2009; Boiy and Moens, 2009; Maks and Vossen, 2012; Desmet and Hoste, 2013; Yu et al., 2013; Abdul-Mageed et al., 2014; Chenlo and Losada, 2014; Appel et al., 2016). Sentiment Analysis at sentence level aims to classify opinions in each sentence. It consists of two classification tasks (Liu and Zhang, 2012). First, subjectivity classification aims to distinguish sentences that reflect information and facts (sentences objectivity) from sentences that express views and opinions (subjective sentences). The second one is polarity classification of the sentences (positive or negative), which we will discuss in detail in Section 2.3.

2.2.3 Aspect-level

The classification of text sentiments at document and sentence level is essential in many cases, but it does not present all the required details. For example, being positive about a particular entity does not imply that the author's opinion is positive about all the aspects of an entity. Similarly, negative sentiments do not represent the negative author opinion about all the aspects of an entity (Liu and Zhang, 2012). For fine-grained comparison of two or more products of similar categories, we need to figure out pros and cons of various components and features (aspects). Classification at document or sentence level does not afford these type of information, and we need to perform opinion mining at aspect level to gain these details. Aspect level opinion mining examines the given opinion itself instead of looking to the language structures

(document, sentence or phrase) (Liu and Zhang, 2012). The purpose of this level is to identify and extract the aspects from each named entity occurring in the text and then assign them a polarity. A summary of the sentiments about different aspects of the desired entity is the most common output derived from this level of sentiment analysis. Many studies have dealt with this level (Miao et al., 2009; Wang et al., 2011; Zhu et al., 2011; Liu et al., 2012; Zhai et al., 2012; Li et al., 2012; Li and Tsai, 2013; Garcia-Moya et al., 2013; Penalver-Martinez et al., 2014; Quan and Ren, 2014). It is worth noticing that this level of opinion mining provides a deeper analysis of the target entity.

2.2.4 Concept-level

The objective of concept-level sentiment analysis is to go beyond a mere word-level analysis of the text and provide new approaches to opinion mining and sentiment analysis that enable a more efficient passage from (unstructured) textual information to (structured) machine-processable data, in any domain. Concept-based methods to sentiment analysis focus on a semantic analysis of the text using web ontologies or semantic networks, which allow the aggregation of conceptual and affective information associated with natural language opinions. By relying on broad semantic knowledge bases, such approaches step away from the modest use of keywords and word co-occurrence counts. Instead of count-based co-occurrences, they rely on the implicit features associated with natural language concepts. Unlike purely syntactical techniques, concept-based approaches can also discover sentiments that are expressed subtly, e.g., throughout the analysis of ideas that do not explicitly convey any emotion, but which are implicitly linked to other concepts that do so (Cambria, 2013).

Cambria et al. (2013) presented the concept level of opinion mining as a new avenue in Sentiment Analysis. The analysis of emotions at concept level is based on the inference of conceptual information about emotion and sentiment associated with natural language. Poria et al. (2014) improved the accuracy of polarity detection through a new approach. An analysis of comments at conceptual level has been proposed that integrates linguistic, common-sense computing, and machine learning techniques. Their results show that the proposed method has a desirable accuracy and better than conventional statistical methods. A concept level sentiment dictionary has been built by Tsai et al. (2013) based on common-sense knowledge. There are also several studies that focused on sentiment analysis at the conceptual level such as Poria et al. (2013); Balahur et al. (2012); Cambria et al. (2015); Weichselbraun et al. (2013); Shah et al. (2016).

2.3 Sentiment Classification Methods

The most important and critical step of opinion mining is selecting an appropriate technique to classify the sentiments.

Sentiment classification, also termed as polarity determination, is concerned with determining the polarity of an object (document, sentence, etc.), whether an object is expressing positive, negative or neutral sentiment towards the subject. As such, it has been applied to social media networks, product reviews, forums, blogs, news articles, and so on.

In this section, we explain, categorize, summarize and compare proposed techniques in this area. The classification methods which are offered in the literature can fall into three groups: machine learning, lexicon-based, and hybrid approaches.

Machine learning approaches can be categorized into two main categories: supervised and unsupervised techniques. The success of both is mainly based on the selection and extraction of the appropriate set of features used to detect sentiments. NLP techniques play a critical role in providing the features. Some of the most important features used are : (1) terms (words or n-grams) and their frequency; (2) part-of-speech information, since adjectives and adverbs play an essential role (Benamara et al., 2007), while nouns can also provide polarity information (Taboada et al., 2011); (3) negations change the meaning of any sentence; (4) syntactic dependencies (tree parsing) can determine the meaning of punishment; among others (Liu and Zhang, 2012; Chenlo and Losada, 2014; Serrano-Guerrero et al., 2015). In Chapter 4, we will discuss the main linguistic features proposed by some relevant studies.

Lexicon-based techniques classify a document/sentence/aspect as positive or negative based on lists of words that represent the two basic polarity classes, such as *good, wonderful, beautiful, amazing,.., etc* or *bad, awful, ugly, terrible,.., etc*.

The hybrid approach combines two or more techniques such as lexicon approach, supervised machine learning, unsupervised machine learning, or even all of them together to improve the sentiment classification performance.

Although all these approaches are categorized separately, they are intertwined with each other, for example, supervised machine learning can be trained with linguistic features derived from polarity lexicons, while the classification algorithms underlying most unsupervised methods are mainly based on the use of polarity lexicons. So, most approaches might be considered as hybrid. Notice that, even if supervised and unsupervised strategies tend to be called machine learning approaches, some unsupervised techniques, namely those using po-

larity lists of words, are often known as lexicon-based. As shown in Fig. 2.3, these approaches are used integrally to obtain a more precise sentiment classification.

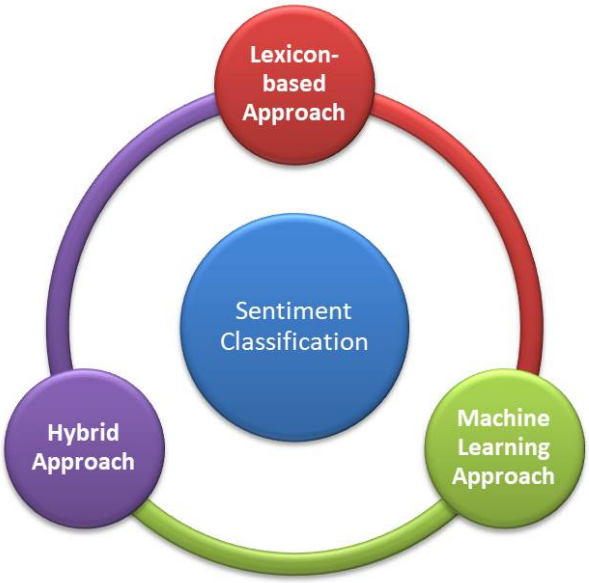

Figure 2.3: Sentiment Classification approaches.

2.3.1 Machine Learning Approaches

2.3.1.1 Supervised Machine Learning Methods

Supervised learning approaches use labeled training documents based on automatic text classification. A labeled training set with a pre-defined category is applied. A classification model is built to predict the class of document based on pre-defined types. Fig. 2.4 shows types of classifiers for sentiment classification of supervised learning algorithms:

- Probabilistic classifiers like Naive Bayes, Bayesian network, and maximum entropy.
- Decision tree classifiers build a hierarchical tree-like structure with true/false queries based on categorization of training document.

– Rule-based classifiers divide the data into a set of rule. The rule in the form of "IF condition THEN conclusion" is generated during the training phase. Decision rules classification method classifies documents to annotated categories.

– Linear classifiers determine good separators with can best separate the space into different classes. Most famous linear classifiers are SVM and neural networks(NN).

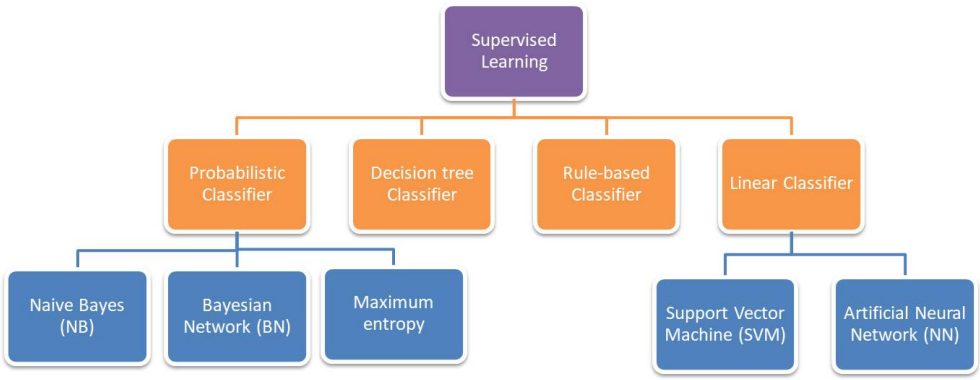

Figure 2.4: Supervised sentiment classifiers.

The most famous and pioneer research on document-level sentiment analysis was conducted by Pang et al. (2002) using Naive Bayes (NB), Maximum Entropy (ME), and SVM for binary sentiment classification of movie reviews. They also tested different features, to find out that SVM with unigrams yielded the highest accuracy. Another finding was that discourse analysis, focus detection, and co-reference resolution could improve the accuracy.

SVM is one of the most popular supervised classification methods. It has a robust theoretical base, is likely the most precise method in text classification (Liu, 2007) and is also successful in sentiment classification (Mullen and Collier, 2004; Saleh et al., 2011; Kranjc et al., 2015). It generally outperforms Naive Bayes and finds the optimal hyperplane to divide classes (Joachims, 1998). Moraes et al. (2013) compared SVM and NB with Artificial Neural Network (NN) approaches for sentiment classification. Experiments were performed on the

both balanced and unbalanced dataset. For this purpose, four datasets were chosen, movies review dataset (Pang and Lee, 2004a) and three different products review: GPS, Books, and Cameras. For unbalanced dataset, the performances of both classifiers, NN and SVM, were affected. Bilal et al. (2016) compared the efficiency of three techniques, namely Naive Bayes, Decision Tree, and Nearest Neighbour in order to classify Urdu and English opinions in a blog. Their results show that Naive Bayes has better performance than the other two. Table 2.1 summarizes the main components of some published studies: techniques utilized, the granularity of the analysis (sentence-level or document-level, etc.), type of data, source of data, and language.

As we focus on demonstrating the efficiency of linguistic features together with the sentiment lexicons we have built, regardless of the classification approach itself, we need a classifier that has already been proved successful with text classification in general and sentiment classification in particular. Since SVM is the best for the text classification, we used it as a significant classifier in all the experiments we conducted with supervised learning in this thesis, either with linguistic features or with sentiment lexicons. We are not comparing the classifiers with each other, and that is why we use the SVM.

Ref.	Techniques Utilized	Granularity	Type of Data	Language
Pang et al. (2002)	ME,NB,SVM	Document level	Movie reviews	English
Boiy and Moens (2009)	MNB[1], ME, SVM	Sentence level	Blog, Review and News forum	English, Dutch, French
Saleh et al. (2011)	SVM	Document level	Movie, Hotel, Products	English
Xia et al. (2011)	NB, ME, SVM	Document level	Movie, Products	English
Abbasi et al. (2011)	Rule-based, SVM	Sentence level	Movie, Products	English
Moraes et al. (2013)	SVM, NN	Document level	Movie, GPS, products	English
Duwairi and Qarqaz (2014)	NB, SVM, KNN	Document level	Education, sports, political news	Arabic
Habernal et al. (2015)	ME, SVM	Document level	Movie, Products	Czech
Jeyapriya and Selvi (2015)	2015	NB	Sentence level	Products English
Severyn et al. (2016)	SVM	Document level	Products	English, Italian
Pham and Le (2018)	NN	Aspect level	Hotel	English

Table 2.1: Main components of some supervised learning sentiment classification published studies.

2.3.1.2 Unsupervised Machine Learning Methods

Unlike supervised learning, unsupervised learning approaches do not depend on the domain and topic of training data. Unsupervised learning overcomes the difficulty of collecting and creating labeled training data.

In unsupervised learning methods, a set of training samples is considered for which only the input value is specified, and the accurate information about the output is not available.

Unsupervised machine learning does not require a big amount of human-annotated training data to obtain acceptable results. This has motivated us to look for methods that do not need training data or need only a relatively small amount of it. The most popular unsupervised classification strategies used in sentiment analysis are classification using syntactic patterns and lexical-based methods (Liu, 2015). Lexical-based methods are seen as a particular instance of unsupervised approaches, since they consists of a predefined list of words associated with a specific sentiment. in Chapter 6 we will explain an unsupervised approach to search for extreme opinions, which is based on the automatic construction of a new lexicon containing the most negative and most positive words that described in Chapter 3.

A simple unsupervised learning algorithm was presented by Turney (2002). He classifies reviews into two categories, *recommend* or *not recommend*, depending on the average number of positive and negative phrases appearing in the review. His algorithm consists of the following steps: first, it searches for phrases in the review by using a part-of-speech (POS) tagger (see Table 2.2). Two consecutive words are extracted if their POS tags conform to any of the patterns in Table 2.3. The polarity of the extracted phrases is then determined by computing Pointwise Mutual Information and Information Retrieval (PMI-IR). Next, the algorithm identifies those associative words returned by the search engine using the NEAR operator. Finally, the polarity of each phrase is solved by computing all the polarities returned by the search engine.

CC	conjunction, coordinating	PRP$	pronoun, possessive
CD	cardinal number	RB	adverb
DT	determiner	RBR	adverb, comparative
EX	existential there	RBS	adverb, superlative
FW	foreign word	RP	adverb, particle
IN	conjunction, subordinating or preposition	SYM	symbol
JJ	adjective	TO	infinitival to
JJR	adjective, comparative	UH	interjection
JJS	adjective, superlative	VB	verb, base form
LS	list item marker	VBZ	verb, 3rd person singular present
MD	verb, modal auxillary	VBP	verb, non-3rd person singular present
NN	noun, singular or mass	VBD	verb, past tense
NNS	noun, plural	VBN	verb, past participle
NNP	noun, proper singular	VBG	verb, gerund or present participle
NNPS	noun, proper plural	WDT	wh-determiner
PDT	predeterminer	WP	wh-pronoun, personal
POS	possessive ending	WP$	wh-pronoun, possessive
PRP	pronoun, personal	WRB	wh-adverb

Table 2.2: Penn Treebank part-of-speech (POS) tags.

First Word	Second Word	Third Word
JJ	NN or NNS	Anything
RB,RBR, or RBS	JJ	not NN nor NNS
JJ	JJ	not NN nor NNS
NN or NNS	JJ	not NN nor NNS
RB,RBR, or RBS	VB,VBD,VBN, or VBG	Anything

Table 2.3: Patterns of POS by Turney Turney (2002)

2.3.2 Lexicons-based Approaches

Sentiment words, also called opinion words, are considered the primary building block in sentiment analysis as it is an essential resource for most algorithms, and the first indicator to express positive or negative opinions. For instance, *good, great, amazing,* and *wonderful* are positive, where as *bad, awful, poor* and *terrible* are negative. In addition to individual words, there are also phrases, which consist of more than one word and express positive or negative opinions, e.g.*very good*.

A list of such words and phrases is called a sentiment lexicon (or opinion lexicon) (Pang et al., 2008).

Some of the lexicons were created based on the part-of-speech (POS), where the words contained in the dictionary were divided into adjectives, adverbs, nouns, and verbs Turney (2002).

Ding et al. (2008) proposed a holistic lexicon-based approach which improved the lexicon-based method proposed by Hu and Liu (2004). Their approach solved the context-dependent problem of opinion words by utilizing information from other sentences rather than looking at only one sentence. This strategy takes some linguistic properties of natural language expressions into account in order to infer the polarity of opinion words. It requires no prior domain knowledge or user inputs. The authors also proposed a solution to the problem of having multiple conflicting opinion words in a sentence, by considering the distance between each opinion word and the product feature by considering the distance between each opinion word and the product feature (rather than the whole review).

Other research conducted by Takamura et al. (2007) suggests a method for extracting polarity for phrases. They build lexical networks connecting similar words with two types of links: words linked with the same polarity and those with different polarity. The proposed method can classify adjective-noun phrases consisting of unseen words. Mohammad and Turney (2013a) built a lexicon containing a combination of sentiment polarity (positive, negative) with one of eight possible emotion classes (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) for each word. Lin et al. (2013) suggested a cross-language opinion lexicon extraction framework using the mutual-reinforcement label propagation algorithm. Zhang and Singh (2014) proposed a semi-supervised framework for generating a domain-specific sentiment lexicon to reduce the human effort for constructing a high-quality domain-specific sentiment lexicon.

However, the most influential words in sentiment analysis are adjectives and adjective phrases, as they are the main origin of subjective content in any document. This is the reason why most early research focused on the use of qualities (Hatzivassiloglou and McKeown, 1997; Hu and Liu, 2004; Taboada et al., 2006). In Section 2.4, we will review in details all approaches to build sentiment lexicons.

2.3.3 Hybrid Approaches

Prabowo and Thelwall (2009) developed various hybrid classifiers over five classifiers: (a) General Inquirer based classifier (GIBC) (Stone et al., 1966), (b) rule-based classifier (RBC), (c) statistics based classifier (SBC), (d) induction rule-based classifier (IRBC), and (e) SVM.

Carrillo de Albornoz et al. (2010) presented a hybrid approach based on machine learning techniques and lexical rules to classify the polarity and intensity of sentences. Their method can determine the polarity of each sentence (negative or positive), as well as its intensity. The system looks over the effect of negations and quantifiers in sentiment analysis and addresses the problem of word ambiguity. Ghiassi et al. (2013) introduced a hybrid approach that uses n-gram analysis for feature extraction and a dynamic artificial neural network (DAN2) (Ghiassi and Saidane, 2005) algorithm or SVM as alternative approaches for Twitter sentiment analysis. Poria et al. (2014) introduced a hybrid approach comprises linguistics, commonsense computing, and machine learning for concept-level sentiment analysis. Appel et al. (2016) proposed a hybrid approach at sentence level, using semantic rules, improved negation management, and an enhanced sentiment lexicon to identify sentiment polarity. They also computede the intensity of sentiment polarity using fuzzy sets as a fundamental tool.

2.4 Sentiment Lexicon construction

There are three main ways of building sentiment lexicons: hand-craft elaboration, automatic expansion from an initial list of seed words and corpus-based approaches. Corpus-based approaches also make use of a list of seed sentiment words to find other sentiment words and their polarity from the given corpus (see Fig. 2.5).

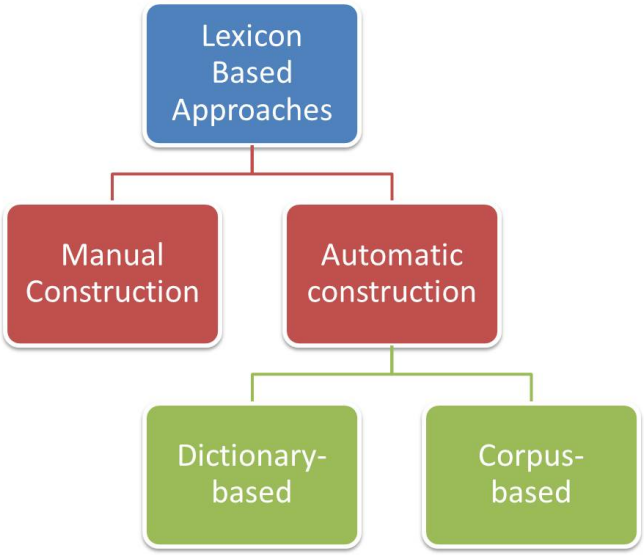

Figure 2.5: Overview of Sentiment Lexicon construction approaches.

2.4.1 Manual Constructed Sentiment Lexicons

Traditionally, in lexicography, lexical resources are made entirely by hand.This is a traditional method of lexicon construction where the lexicons are built by hand (human-based). Manually creating a comprehensive sentiment lexicon is an intensive labor and sometimes error prone process, so it is no wonder that many opinion mining researchers and practitioners rely so heavily on existing lexicons as primary resources. According to Cruz et al. (2014); Joshi et al. (2017), General Inquirer (Stone et al., 1966) can be considered the first hand-made sentiment lexicon. It is a dictionary constituted by semantic units that can appear in multiple lexicalized forms, called lemmas, e.g., the verb "talk" is a lemma that can be found in texts with different inflections, like "talked" or "talking"; it also includes a significant amount of information (syntactic, semantic and pragmatic) related to each lemma, and it has 4,206 lemmas tagged as positive or negative.

General Inquirer is still widely used in many works on Sentiment Analysis, in spite of its age.

Nielsen et al. (2011) has presented another manually generated lexicon called AFINN. In this lexicon, a list of English words has been constructed and rated for valence with an integer

between minus five (negative) and plus five (positive). Taboada et al. (2011) created their dictionary SO-CAL manually since they believe that the overall accuracy of lexicon-based sentiment analysis mainly relies on the quality of those resources. The lexicon was built with content words, namely adjectives, adverbs, nouns, and verbs, adding sentiment scores between -5 and +5. The negative sign (-) refers to negative polarity, positive sign (+) indicates positive polarity, while any semantically neutral word has zero score.

Hutto and Gilbert (2014) offered a new handcraft lexicon. Over 7500 token features were rated on a scale from –4, Extremely Negative, to +4, Extremely Positive, with allowance for 0, meaning Neutral.

The NRC emotion lexicon is a list of words and their associations with emotions and sentiments (negative and positive). The annotations were manually done through Amazon's Mechanical Turk (Mohammad and Turney, 2010, 2013b).

2.4.2 Automatic Construction of Sentiment Lexicons

The two automatic techniques for creating polarity lexicons are dictionary and corpus-based methods, which we will describe in the following subsections.

2.4.2.1 Dictionary-based

This strategy requires seed sentiment words to bootstrap new polarity entries. They are mainly based on the synonyms and antonyms structure of external resources, such as thesaurus.

Kamps et al. (2004) report a thesaurus-based method that makes use of the synonymy relation between adjectives in WordNet (Fellbaum, 1998) to generate a graph. More precisely, the authors measure the shortest path between the adjective and two basic sentiment seeds, "good" and "bad", to determine the polarity of a word. This is a semi-supervised learning method which starts with a lexical resource, WordNet, and a small list of polarity words in order to expand the lexical resource in an iterative process. In a similar way, Kim and Hovy (2006) propose a method that starts with three seed lists containing positive, negative and neutral words, which are also expanded with their synonyms in WordNet.

Unlike these strategies, our method does not require any thesaurus to expand the lexicon with synonyms or antonyms.

2.4.2.2 Corpus-based

The main concern of lexicon-based approaches is that most polarity words are domain dependent since the subjectivity status of most words is very ambiguous. The same word may be provided with a subjective burden in a specific domain while it can refer to objective information in another area. It follows that domain dependent lexicons should outperform general-purpose dictionaries in the task of sentiment analysis. However, the construction of domain-dependent polarity lexicons is a strenuous and tedious task if it is made manually for each target domain. With the increasing of many sentiment corpora in diverse domains, the automatic generation of this kind of resources for many domains is becoming a fundamental task in opinion mining and sentiment analysis (Huang et al., 2014). The corpus-based techniques arise with the objective of providing dictionaries related to a specific domain.

Lu et al. (2011) present an automated approach for constructing a context-dependent lexicon from an unlabeled opinionated text collection based on existing lexicons and tagged consumer reviews. Each entry of this lexicon is a pair containing a sentiment term and different "aspect" terms associated with the former. The same sentiment term may diverge in polarity when co-occurring with a particular aspect term. This strategy is semi-supervised since it needs to start with a seed list of words or with an existing lexicon. Molina-González et al. (2015) generated a polarity word list by integrating most frequently used positive and negative words in eight different domain reviews on Spanish datasets only. Liao et al. (2016) proposed a hybrid method of domain lexicon construction which explores syntactic and semantic information through part-of-speech, dependency structure, phrase structure, semantic role and semantic similarity.

By contrast, our method generates the lexicon of positive and negative adjectives and adverbs directly from any labeled corpus for any language without needs to start with the small set of words as a seed or any existing lexicon.

Table 2.4 summarizes the main components of some lexicon-based studies: construction method, granularity of the analysis (sentence-level or document-level,..,etc.), type of data, source of data, POS (apply part-of-speech (yes/no)) and language.

The previous related works did not focus on the most negative and most positive words. We propose a new method to build opinion lexicons from multiple domains for the most negative and most positive words, which is quite a different resource with regard to existing lexicons. As far as we know, no previous work has been focused on detecting extreme opin-

Ref.	Construction Method	Granularity	Type of Data	POS	Language
Nasukawa and Yi (2003)	Manually	Sentence level	Web pages, News	Yes	English
Kanayama and Nasukawa (2006)	Corpus-based	Sentence level	Movie, Products	Yes	Japanese
Taboada et al. (2011)	Manually	Sentence level	Movie, Products	Yes	English
Sharma et al. (2014)	Dictionary based	Document level	Movie reviews	Yes	English
Chinsha and Joseph (2015)	Dictionary based	Aspect level	Restaurant reviews	Yes	English
Jiménez-Zafra et al. (2016)	Dictionary and Corpus based	Aspect level	Products, Restaurant	-	English Spanish
Rathan et al. (2018)	Dictionary based	Aspect level	Mobile reviews	Yes	English
Chao and Yang (2018)	Corpus-based	Aspect level	Restaurant reviews	Yes	Chinese

Table 2.4: Main components of some lexicon-based published studies.

ions. Our proposal, which we will describe in Chapter 3, therefore, may be considered to be a first step in that direction.

2.5 Evaluation Methodology: Lexicons and Datasets

In this section, we review the most popular sentiment lexicons that have been built, in addition to datasets, which are used as a benchmark to evaluate the performance of sentiment analysis systems.

2.5.1 Lexicons

2.5.1.1 Hu & Liu Opinion Lexicon

Hu & Liu Opinion Lexicon[2] is a list of 6790 negative and positive words for English: 2007 positive, and 4783 negative words. This list was accumulated across several years starting from Hu and Liu (2004); Liu et al. (2005). It includes mis-spellings, morphological variants, slang, and social-media mark-up.

In contrast to our dictionaries that we have built upon our proposed methodology in Chapter 3, this lexicon did not rank the words according to their strength scale; they merely put two lists of words either positive or negative, which makes it unfit to be used as a feature to identify the extreme reviews.

[2] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

2.5.1.2 SO-CAL

Sentiment Orientation Calculator (SO-CAL) was described in Taboada et al. (2011). The authors created their dictionary manually by a native English speaker and reviewed by a committee of three researchers since they believe that the overall accuracy of lexicon-based sentiment analysis mainly relies on the quality of those resources. The lexicon was built with content words, namely adjectives, adverbs, nouns, and verbs. The enhanced dictionaries contain 2826 adjective entries, 1549 nouns, 1142 verbs, and 876 adverbs.

Adjectives were taken from a corpus of 400 Epinion[3] reviews extracted from eight different categories: books, cars, computers, cookware, hotels, movies, music, and phones. In addition to Epinions reviews, the separated noun, verb, and adverb dictionaries were also taken from:

– A subset of 100 movie reviews from the Polarity Dataset Pang and Lee (2004b).

– Positive and negative words from General Inquirer.

Each term is assigned a sentiment valence value on a scale of -5 to +5 (neutral or zero-value words were excluded). Compared to the other words in the dictionary, there is a few number of extreme words. In adjectives dictionary, 189 words were ranked as -4 and -5; on the other hand 239 were rated as +4 and +5. So, only 0.06% are very negative and 0.08% very positive words, which means that the number of extreme words is very few compared to other words with a non-extreme value on the scale. The same situation happens in the adverbs dictionary. There are only 62 adverbs rated as very negative (only 0.07%) and 75 as very positive (0.08%).

As very negative and very positive words are appropriate as evidence for searching for extreme views, the method described in Chapter 3 will focus on the automatic identification of this kind of polarity words.

2.5.1.3 MPQA Subjectivity Lexicon

The MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon[4] is maintained by Wilson et al. (2005). The clues in this lexicon were gathered from many sources. Some were gleaned from manually developed resources, while others were classified automatically using

[3] `www.epinions.com` (now offline)

[4] `http://mpqa.cs.pitt.edu/lexicons/`

both annotated and unannotated data. Most of the clues were collected as part of the work reported by Riloff and Wiebe (2003). This lexicon contains a list of 8222 single words: 4911 negative, 2718 positive, 570 neutral and 21 words are both positive and negative. Statements concerning each word such as strength, part-of-speech and polarity are shown in Fig. 2.6

Strength	Length	Word	POS	Stemmed	Polarity
type=weaksubj	len=1	word1=back	pos1=verb	stemmed1=y	priorpolarity=positive
type=strongsubj	len=1	word1=backbite	pos1=verb	stemmed1=y	priorpolarity=negative
type=strongsubj	len=1	word1=backbiting	pos1=anypos	stemmed1=n	priorpolarity=negative
type=weaksubj	len=1	word1=backbone	pos1=noun	stemmed1=n	priorpolarity=positive
type=strongsubj	len=1	word1=backward	pos1=adj	stemmed1=n	priorpolarity=negative
type=strongsubj	len=1	word1=backwardness	pos1=noun	stemmed1=n	priorpolarity=negative
type=strongsubj	len=1	word1=bad	pos1=adj	stemmed1=n	priorpolarity=negative
type=strongsubj	len=1	word1=badly	pos1=adj	stemmed1=n	priorpolarity=negative
type=strongsubj	len=1	word1=baffle	pos1=verb	stemmed1=y	priorpolarity=negative
type=strongsubj	len=1	word1=baffled	pos1=adj	stemmed1=n	priorpolarity=negative
type=strongsubj	len=1	word1=bafflement	pos1=noun	stemmed1=n	priorpolarity=negative
type=strongsubj	len=1	word1=baffling	pos1=anypos	stemmed1=n	priorpolarity=negative
type=strongsubj	len=1	word1=bait	pos1=verb	stemmed1=y	priorpolarity=negative

Figure 2.6: A piece of the MPQA subjectivity lexicon.

2.5.1.4 AFINN-111

Nielsen (2011) presented another manually generated lexicon called AFINN[5]. In this lexicon, a list of English words was constructed and rated for valence with an integer between -5 (negative) and +5 (positive). The words have been manually labeled. AFINN-111 has 2477 words and phrases: 1598 negative, 878 positive and 1 neutral words.

2.5.1.5 SentiWordNet

SentiWordNet[6] is a lexical resource for opinion mining described in details by Esuli and Sebastiani (2007); Baccianella et al. (2010). SentiWordNet assigned to each synset of WordNet three sentiment scores: positivity, negativity, objectivity as follows: Pos(s): the positive score of synset s Neg(s): The negative score of synset s Obj(s): The objective score of synset s where,

$$0 \leq Pos(s), Neg(s), Obj(s) \leq 1$$

[5] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
[6] http://sentiwordnet.isti.cnr.it

$$Pos(s) + Neg(s) + Obj(s) = 1$$

For instance, the scores for the synset beautiful#1 are:

– Pos(beautiful#1) = 0.75

– Neg(beautiful#1) = 0.00

– Obj(beautiful#1) = 0.25

This formulation of the sentiment values has the following salient features:

– The sentiment is tied intimately to the meaning of a word rather than the word itself.

– A synset is allowed to be both positive and negative, or neither positive nor negative.

– The sentiment evaluation is ranked over a scale instead of a binary or ternary classification.

Fig. 2.7 shows a sample of the components of a SentiWordNet lexicon: part-of-speech(POS), ID, positive score (PosSore), negative score (NegScore), SynsetTerms, and glossary.

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	207034	0	0	rostrate#1	having a beak or beaklike snout or proboscis
a	207133	0	0	short-billed#1 short-beaked#1	having a short beak
a	207226	0.25	0	stout-billed#1	having a strong beak
a	207305	0	0	straight-billed#1	having a straight beak
a	207389	0	0	thick-billed#1	having a thick beak
a	207467	0	0.625	beakless#1	not having a beak or bill
a	207547	0	0	bedded#2	having a bed or beds as specified
a	207706	0.125	0	double-bedded#1	having a double bed; "a double-bedded room"
a	207809	0	0	single-bedded#1	having single beds
a	207887	0	0	twin-bedded#1	having twin beds
a	207961	0	0.5	bedless#1	without a bed; "the cell was bedless"
a	208052	0	0	beneficed#1	having a benefice; "a beneficed clergyman"
a	208150	0	0.625	unbeneficed#1	not having a benefice

Figure 2.7: A fragment of the SentiWordNet lexicon.

2.5.1.6 SentiWords

SentiWords is a sentiment lexicon derived from SentiWordNet using the method described in Gatti et al. (2016). It contains more than 16,000 words associated with a sentiment score between -1 (very negative) and +1 (very positive). The words in this lexicon are arranged with WordNet synsets, that include adjectives, nouns, verbs and adverbs.

2.5.1.7 SentiStrength

SentiStrength[7] estimates the strength of positive and negative sentiment in short texts, even for informal language. It has human-level accuracy for short social web texts in English, except political texts. SentiStrength can report binary (positive/negative), trinary (positive/negative/neutral) and single scale (-4 to +4) results. SentiStrength was originally developed for English and optimized for general short social web texts but can be configured for other languages and contexts by changing its input files (Thelwall et al., 2010).

2.5.1.8 SenticNet

SenticNet[8] is a public word source made by Sentic computing. The polarity score for each concept is calculated in $[1,-1]$ interval. There are five versions of SenticNet and all of them are available in RDF/XML format (see Fig. 2.8).

SenticNet 1 (Cambria et al., 2010) simply associated polarity scores with almost 6,000 ConceptNet concepts; in addition to polarity, SenticNet 2 (Cambria et al., 2012) also assigned semantics and sentics to commonsense concepts and extended the breadth of the knowledge base to about 13,000 entries; SenticNet 3 (Cambria et al., 2014) broadened the spectrum of the semantic network to 30,000 concepts; SenticNet 4 (Cambria et al., 2016) introduced the concept of semantic primitives to further extended the knowledge base to 50,000 entries; finally, SenticNet 5 (Cambria et al., 2018) reaches 100,000 commonsense concepts by employing recurrent neural networks to infer primitives by lexical substitution.

[7] http://sentistrength.wlv.ac.uk
[8] https://sentic.net/downloads/

```
<rdf:Description rdf:about="http://openmind.media.mit.edu/api/en/concept/taste">
        <rdf:type rdf:resource="http://sentic.net/api/concept"/>
        <text xmlns="http://sentic.net/api/">taste</text>
        <polarity xmlns="http://sentic.net/api/"
rdf:datatype="http://www.w3.org/2001/XMLSchema#float">+0.197</polarity>
    </rdf:Description>

    <rdf:Description
rdf:about="http://openmind.media.mit.edu/api/en/concept/taste%20bad">
        <rdf:type rdf:resource="http://sentic.net/api/concept"/>
        <text xmlns="http://sentic.net/api/">taste bad</text>
        <polarity xmlns="http://sentic.net/api/"
rdf:datatype="http://www.w3.org/2001/XMLSchema#float">-0.148</polarity>
    </rdf:Description>
```

Figure 2.8: A piece of SenticNet 3 Lexicon.

2.5.1.9 VADER

Valence Aware Dictionary and Sentiment Reasoner (VADER) are respectively a lexicon and and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media working well on texts from other domains (Hutto and Gilbert, 2014). The authors obtained over 7,500 lexical features with validated valence scores indicating both sentiment polarity (negative/positive) and sentiment intensity from 10 independent human raters on a scale from –4, Extremely Negative, to +4, Extremely Positive, with allowance for 0, Neutral. Only the lexical feature that had a non-zero mean rating and standard deviation less than 2.5 as determined by the aggregate of the ten independent raters were kept (see Fig. 2.9).

TOKEN	MEAN-SENTIMENT-RATING	STANDARD DEVIATION	RAW-HUMAN-SENTIMENT-RATINGS
activeness	0.6	0.8	[0, 2, 0, 0, 1, 0, 1, 0, 2, 0]
activenesses	0.8	0.74833	[2, 0, 1, 0, 0, 0, 1, 2, 1, 1]
actives	1.1	0.7	[2, 1, 0, 1, 1, 0, 1, 1, 2, 2]
adequate	0.9	0.7	[0, 0, 1, 1, 0, 2, 1, 1, 2, 1]
admirability	2.4	0.4899	[2, 3, 3, 3, 3, 2, 2, 2, 2, 2]
admirable	2.6	0.66332	[2, 3, 3, 3, 4, 3, 2, 2, 2, 2]
admirableness	2.2	0.87178	[2, 2, 3, 3, 3, 1, 3, 1, 3, 1]
admirably	2.5	0.67082	[2, 3, 3, 3, 4, 2, 2, 2, 2, 2]
admiral	1.3	1.18743	[0, 0, 1, 3, 3, 2, 2, 0, 2, 0]
admirals	1.5	0.80623	[2, 2, 0, 2, 2, 0, 1, 2, 2, 2]
admiralties	1.6	0.66332	[2, 2, 2, 1, 0, 2, 2, 2, 1, 2]
admiralty	1.2	1.53623	[0, 4, 0, 0, 0, 2, 2, 3, 2, -1]
admiration	2.5	0.80623	[3, 1, 1, 3, 3, 2, 3, 3, 3, 3]
admirations	1.6	0.66332	[2, 2, 1, 1, 2, 2, 2, 2, 2, 0]
admire	2.1	0.83066	[3, 3, 1, 3, 3, 2, 1, 2, 1, 2]
admired	2.3	0.78102	[4, 2, 2, 2, 2, 2, 3, 3, 1, 2]
admirer	1.8	0.74833	[2, 1, 1, 2, 3, 2, 3, 1, 1, 2]
admirers	1.7	1.00499	[2, 3, 2, 2, 2, 1, -1, 2, 2, 2]
admires	1.5	0.67082	[3, 1, 1, 2, 1, 2, 2, 1, 1, 1]
admiring	1.6	0.8	[1, 2, 1, 1, 3, 3, 2, 1, 1, 1]

Figure 2.9: A piece of the Valence Aware Dictionary and Sentiment Reasoner (VADER) Lexicon.

2.5.1.10 LIWC

Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) is a popular manually created emotion lexicon. It consists of 290 words and word-stems. Each word or word-stem defines one or more word categories or sub-dictionaries. This type of emotion lexicons such as LIWC, ANEW, and NRC that we will describe next, are not considered for our work since they are not polarity lexicons.

2.5.1.11 ANEW

The ANEW[9] (Affective Norms for English Words) provides a set of normative emotional ratings for a large number of words in the English language. This set of verbal materials has been rated regarding pleasure, arousal, and dominance to create a standard for use in studies of emotion and attention (Bradley and Lang, 1999) ANEW words have been ranked regarding their pleasure, arousal, and dominance. ANEW words have an associated sentiment valence ranging from 1-9 (with a neutral midpoint at five), such that words with valence scores

[9] http://csea.phhp.ufl.edu/media/anewmessage.htm

less than five are considered negative, and those with scores higher than five are considered positive (see Fig. 2.10).

Description	Word No.	Valence Mean(SD)	Arousal Mean(SD)	Dominance Mean (SD)	Word Frequency
honest	210	7.70 (1.43)	5.32 (1.92)	6.24 (2.13)	47
honey	792	6.73 (1.70)	4.51 (2.25)	5.44 (1.47)	25
honor	211	7.66 (1.24)	5.90 (1.83)	6.70 (2.04)	66
hooker	793	3.34 (2.31)	4.93 (2.82)	4.73 (2.48)	.
hope	794	7.05 (1.96)	5.44 (2.47)	5.52 (2.20)	178
hopeful	212	7.10 (1.46)	5.78 (2.09)	5.41 (1.92)	12
horror	213	2.76 (2.25)	7.21 (2.14)	4.63 (2.70)	17
horse	214	5.89 (1.55)	3.89 (2.17)	4.67 (1.60)	117
hospital	215	5.04 (2.45)	5.98 (2.54)	4.69 (2.16)	110
hostage	216	2.20 (1.80)	6.76 (2.63)	2.83 (2.32)	2
hostile	217	2.73 (1.50)	6.44 (2.28)	4.85 (2.58)	19
hotel	795	6.00 (1.77)	4.80 (2.53)	5.12 (1.84)	126
house	563	7.26 (1.72)	4.56 (2.41)	6.08 (2.12)	591
hug	218	8.00 (1.55)	5.35 (2.76)	5.79 (2.41)	3
humane	796	6.89 (1.70)	4.50 (1.91)	5.70 (1.91)	5
humble	219	5.86 (1.42)	3.74 (2.33)	4.76 (2.25)	18

Figure 2.10: Sample of Affective Norms for English Words (ANEW).

2.5.1.12 NRC

Word-Emotion Association Lexicon (NRC)[10] (also called EmoLex) is a list of 14,182 words and their associations with emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive)Mohammad and Turney (2013b, 2010). This dictionary has been translated into about 40 languages (see Figs. 2.11 and 2.12).

[10] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

English (en)	Positive	Negative	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
backtrack	0	0	0	0	0	0	0	0	0	0
backward	0	1	0	0	0	0	0	0	0	0
backwards	0	1	0	0	1	0	0	0	0	0
backwater	0	1	0	0	0	0	0	1	0	0
bacteria	0	1	0	0	1	1	0	1	0	0
bacterium	0	1	0	0	1	1	0	0	0	0
bad	0	1	1	0	1	1	0	1	0	0
badge	0	0	0	0	0	0	0	0	0	1
badger	0	1	1	0	0	0	0	0	0	0
badly	0	1	0	0	0	0	0	1	0	0
badness	0	1	1	0	1	1	0	0	0	0
baffle	0	0	0	0	0	0	0	0	0	0
bag	0	0	0	0	0	0	0	0	0	0
baggage	0	0	0	0	0	0	0	0	0	0
baggy	0	0	0	0	0	0	0	0	0	0
bagpipes	0	0	0	0	0	0	0	0	0	0
bail	0	0	0	0	0	0	0	0	0	0

Figure 2.11: Sample of the components of Word-Emotion Association Lexicon (NRC).

English (en)	Afrikaans (af)	Albanian (sq)	Amharic (am)	Arabic (ar)	Armenian (hy)	Azeerbaijani (az)	Basque (eu)
aback	uit die veld geslaan	prapa	ተጭኗል	الى الوراء	շեղում	sanki	aback
abacus	abakus	numërator	abacus	طبلية تاج	անբավարարություն	abacus	abako
abandon	verlaat	braktis	ውጣ	تخلى	լքել	tərk et	bertan behera
abandoned	verlate	braktisur	ተትቷል	مهجور	լքված	tərk etdi	abandonatutako
abandonment	verlating	braktisje	ማቋረጥ	التخلي عن	հրաժարվելով	ləğv	abandono
abate	bedaar	i jap fund	አጥፋ	انحسر	քանդել	boşaltın	abate
abatement	vermindering	pakësim	መቦሰበስ	انحسار	նվազեցում	azaldılması	murrizteko
abba	Abba	Abba	abba	أبا	abba	abba	abba
abbot	abt	abat	አቡኝ	رئيس الدير	աբբոթ	abbot	abade
abbreviate	afkort	shkurtoj	አህጽሮት	اختصر	կրճատել	qısaltmaq	laburtu
abbreviation	afkorting	shkurtim	አህጽሮተ ቃል	الاختصار	հապավումը	kısaltma	laburdura
abdomen	buik	abdomen	ሆድ	بطن	որովայնը	qarın	abdominalak

Figure 2.12: Sample of Word-Emotion Association Lexicon (NRC) translated to some languages.

2.5.2 Opinion-Based Datasets

Opinion-based datasets are standard benchmarks that have been constructed over the years to facilitate experimental research in Opinion Mining and Sentiment Analysis. They consist of sentences extracted from the social networks such as (reviews, Twitter, Facebook, etc). in this Section, we summarize in Table 2.5 some of the most popular datasets.

Dataset	Domain	Language	Resource	star rating	Details	Website
Sentiment Polarity Pang and Lee (2004b)	Movie reviews	English	IMDB.com	1 - 10	1000 positive 1000 negative	https://www.cs.cornell.edu/people/pabo/movie-review-data/
Multi-Domain Sentiment Blitzer et al. (2007b)	Product reviews	English	Amazon.com Yelp.com	1 - 5	2000 reviews for each (Kitchen, Books DVDs, and Electronics)	https://www.cs.jhu.edu/ mdredze/datasets/sentiment/index2.html
Large Movie Review Maas et al. (2011)	Movie reviews	English	IMDB.com	1 - 10	50000 reviews	http://ai.stanford.edu/ amaas/data/sentiment/
Jo and Oh (2011)	Product reviews Resturant reviews	English	Amazon.com Yelp.com	1 - 5	24,184 Electronic device reviews 27,458 Restaurant reviews	http://uilab.kaist.ac.kr/research/WSDM11/
HASH Kouloumpis et al. (2011)	Tweets	English	Twitter		31,861 positive, 64,850 negative, 125,859 Neutral tweets	http://demeter.inf.ed.ac.uk
ISIEVE Kouloumpis et al. (2011)	Tweets	English	Twitter		1520 positive, 200 negative, 2295 Neutral tweets	http://demeter.inf.ed.ac.uk
Cruz et al. (2013)	Product reviews	English	Epinion.com	1 - 5	headphones 587, hotels 988, and cars 972 reviews.	http://www.lsi.us.es/_fermin/index.php/Datasets
LABR Aly and Atiya (2013)	Book reviews	Arabic	Goodreads.com	1 - 5	63,000 reviews	http://www.mohamedaly.info/datasets/labr
Hopinion	Hotel and resturans reviews	Spanish	Tripadvisor.com	1 - 5	17934 reviews	http://clic.ub.edu/corpus/es/node/106

Table 2.5: List of some of the publicly available datasets for Sentiment Analysis

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CHAPTER 3

AUTOMATIC CONSTRUCTION OF SENTIMENT LEXICONS

In this chapter, we will describe our proposed method to create lexicons. Our proposed method is capable of building a lexicon for any field or any language, and for standard or extreme opinions; it only requires to exist labeled corpora. In fact, we will describe two very related methods to build sentiment lexicons.

First, we define a strategy to build sentiment lexicons (positive and negative words) from corpora. Particular attention will be paid to the construction of a domain-specific lexicon from a corpus of movie reviews. The lexicon we built using our strategy is called SPLM.

Second, using a similar strategy as the previous one, we build two opinion lexicons from multiple domains for the most negative and most positive words in order to identify extreme opinions. The two lexicons are called VERY-NEG and VERY-POS.

3.1 Construction of Domain-specific Sentiment Lexicons

There exist two main approaches to finding the sentiment polarity at a document or sentence level. First, machine learning techniques based on training corpora annotated with polarity information and, second, strategies based on polarity lexicons. Lexicon-based approaches are very popular in sentiment analysis and opinion mining, and they play a key role in all applications in this field. The main concern of lexicon-based approaches is that most polarity words are domain-dependent since the subjectivity status of most words is very ambiguous. The

same word may be provided with a subjective burden in a specific domain while it can refer to an objective information in another domain. It follows that domain-dependent lexicons should outperform general-purpose dictionaries in the task of sentiment analysis.

However, the construction of domain-dependent polarity lexicons is a strenuous and boring task if it is made manually for each target domain. With the increasing number of sentiment corpora in diverse domains, the automatic generation of this kind of resources for many domains is becoming a fundamental task in opinion mining and sentiment analysis (Huang et al., 2014).

In this Section we propose a method for automatically building polarity lexicons from corpora. More precisely, we focus on the construction of a domain-specific lexicon from a corpus of movie reviews and its use in the task of sentiment analysis. The experiments reported in this section shows that our automatic resource outperforms other manual general-purpose lexicons when they are used as features of a supervised sentiment classifier.

We detail how to construct a lexicon that ranks words from the negative values to positive ones. The lexicon can be generated using any corpus of reviews labeled with star rating: one star (very negative) to N stars (very positive). The category set is the number of stars that can be assigned to the reviews. For instance, we are provided with 10 categories only if each review can be rated from 1 to 10. The first step to create our proposed lexicon is to measure the relative frequency (RF) for every word w in each category c according to equation 3.5:

$$RF_c(w) = \frac{freq(w,c)}{Total_c} \tag{3.1}$$

where c is any category of the star rating, from 1 to N; $freq(w,c)$ is the number of tokens of the target word in c; and $Total_c$ is the total number of word tokens in c.As in our experiments, the corpus was POS tagged; words are actually represented as (Word, Tag) pairs. Besides, we only work with adjectives and adverbs as they are the most relevant part of speech tags in sentiment analysis for any language Benamara et al. (2007).

The second step is to calculate the average of the RF values for two ranges of categories: negative and positive. For this purpose, it is necessary to define two values: first, a borderline value for negative and positive opinions, which might vary according to the specific star rating of the reviews. Second, the number of neutral categories. For example, if the star rating goes from 1 to 10 categories and we set the borderline in 4 with two neutral categories, the negative reviews would be those rated from 1 to 4, while the positive reviews would be from 7 to 10. So the neutral reviews would be those rated from 5 to 6. Given a borderline value, B, the

average of the negative scores, *Avn*, for a word is computed as follows:

$$Avn(w) = \frac{\sum_{c=1}^{B} RF_c(w)}{B} \tag{3.2}$$

On the other hand, given Nt and N where N is the total number of categories, and Nt is the number of neutral categories, the average of positive scores, *Avp*, for each word is computed in equation 3.3:

$$Avp(w) = \frac{\sum_{c=B+Nt}^{N} RF_c(w)}{B} \tag{3.3}$$

In the following step, the negative and positive words are selected by comparing the values of *Avn* with *Avp*. Given a word w, we compute the difference D in equation 3.4 and assign this value to w, which stands for the final *weight* of the word:

$$D(w) = Avp(w) - Avn(w) \tag{3.4}$$

If the value of $D(w)$ is negative, w will be in the class of negative words. If the value of $D(w)$ is positive, w will be in the positive class.

3.1.1 SPLM

Our proposed lexicon was built from the corpus introduced in Potts (2010). The corpus[1] consists of data gathered from the user-supplied reviews at the IMDB. Each of the reviews in this collection has an associated star rating: one star (very negative) to ten stars (very positive). The reviews were tagged using the Stanford Log-Linear Part-Of-Speech Tagger. Then, tags were broken down into the WordNet Tags: *a* (adjective), *n* (noun), *v* (verb), *r* (adverb). Words whose tags were not part of those syntactic categories were filtered out. The list of selected words was then stemmed.

[1] `http://compprag.christopherpotts.net/code-data/imdb-words.csv.zip`

Word	Tag	Category	Count	Total
bad	a	1	122232	25395214
bad	a	2	40491	11755132
bad	a	3	37787	13995838
bad	a	4	33070	14963866
bad	a	5	39205	20390515
bad	a	6	43101	27420036
bad	a	7	46696	40192077
bad	a	8	42228	48723444
bad	a	9	29588	40277743
bad	a	10	51778	73948447

Table 3.1: A sample of the IMDB collection format for the word "bad" as adjective ("a") in each Category (from 1 to 10)

Table 3.1 shows a sample for the adjective "bad", where *Freq* is the total number of tokens of a (Word,Tag) pair in each Category (from rate 1 to 10), while *Total* is the total number of word tokens in each Category. Notice that Total values are constant for all words but they repeated for each one in order to make processing easier.

The next step is to compute *Avn* and *Avp* for each word. By making use of the equations defined above (3.3, 3.2 and 3.4), we obtain the weights assigned to each word-tag pair. It results in a ranked opinion lexicon, which is freely available[2].

3.2 Construction of Extreme Opinions Lexicons

In this section, we describe how to build two lexicons aimed at identifying extreme opinions:

– one that ranks words in the negative scale, from the most negative values to less negative ones,

– and another lexicon in the positive domain, which arranges values from the most positive to the least positive.

The lexicons can be generated using any corpus of reviews labeled with star rating: one star (very negative) to N stars (very positive). The category set is the number of stars that can be assigned to the reviews. For instance, we are provided with 10 categories only if each review can be rated from 1 to 10.

[2] https://github.com/almatarneh/SPLM-Lexicon

In the same way as the previous strategy to build SPLM, the first step to create our proposed lexicons is to measure the relative frequency (RF) for every word w in each category c according to equation 3.5:

$$RF_c(w) = \frac{freq(w,c)}{Total_c} \tag{3.5}$$

where c is any category of the star rating, from 1 to N; $freq(w,c)$ is the number of tokens of the target word in c; and $Total_c$ is the total number of word tokens in c. As in our experiments, the corpus was PoS tagged, words are actually represented as (word, tag) pairs. Besides, we only work with adjectives and adverbs as we described in the strategy for SPLM construction.

The second step is to calculate the average of RF values for two ranges of categories: very negative (VN) *vs* not very negative (NVN), and very positive (VP) *vs* not very positive (NVP). For this purpose, it is necessary to define a borderline value B for extreme opinions, which might vary according to the specific star rating of the reviews. For instance, if the rating goes from 1 to 10, and the borderline value B=2, the VN reviews are considered those rated from 1 to 2, while VP are those rated from 8 to 10. This is similar if the rating goes from 1 to 5 and the borderline is set at 1. In this case, the VN reviews are considered those rated 1, while VP are those rated 5. Given a borderline value, B, the average of the VN scores, $AvVN$, for a word is computed as follows:

$$AvVN(w) = \frac{\sum_{c=1}^{B} RF_c(w)}{B} \tag{3.6}$$

On the other hand, given $R = N - B$, where N is the total number of categories, the average of NVN values, $AvNVN$, for each word is computed in equation 3.7:

$$AvNVN(w) = \frac{\sum_{c=B+1}^{N} RF_c(w)}{R} \tag{3.7}$$

As for the average of VP scores, $AvVP$, for a word, it is computed in equation **??**:

$$AvVP(w) = \frac{\sum_{c=(N+1)-B}^{N} RF_c(w)}{B} \tag{3.8}$$

And the average of NVP values, $AvNVP$, for each word is computed in equation 3.9:

$$AvNVP(w) = \frac{\sum_{c=1}^{N-B} RF_c(w)}{R} \tag{3.9}$$

In the following step, the objective is to assign polarity weights to words and classify them by using four polarity classes: VN, NVN, VP, and NVP. Extreme words (VN and VP) are separated from not extreme words by just comparing the difference between the average values

obtained by the equations defined above: 3.6, 3.7, 3.8, 3.9. With this simple idea, we build two lexicons: one lexicon in the negative scale from VN to NVN, and another lexicon in the positive scale from VP to NVP. So, given a word w, we compute the differences D_{neg} and D_{pos} in equations 3.10 and 3.11, and assign the resulting values to w:

$$D_{neg}(w) = AvNVN(w) - AvVN(w) \tag{3.10}$$

$$D_{pos}(w) = AvNVP(w) - AvVP(w) \tag{3.11}$$

D_{neg} gives a weight to w within the negative scale, while D_{pos} assigns weights in the positive ranking. These two weights are used to classify words in the four above-mentioned categories and thereby build two new polarity lexicons, which we call *VERY-NEG* and *VERY-POS*. Classification is done with the following basic algorithm:

If the value of $D_{neg}(w)$ is negative, w is in VN class. If $D_{neg}(w)$ is positive, w is in NVN.

If the value of $D_{pos}(w)$ is positive, w is in VP class. If $D_{pos}(w)$ is negative, w is in NVP.

VERY-NEG is a lexicon constituted by words classified as VN or NVN, while VERY-POS is another lexicon consisting of words classified as VP or NVP. In both lexicons, words are ranked by means of the weight returned by D_{neg} or D_{pos}.

3.2.1 VERY-NEG and VERY-POS

Our proposed lexicons were built from another text corpora[3] introduced in Potts (2010, 2011). The corpora consist of online reviews collected from IMDB, Goodreads, OpenTable, Amazon/Tripadvisor. Each of the reviews in this collection has an associated star rating: one star (very negative) to ten stars (very positive) in IMDB, and one star (very negative) to five stars (very positive) in the other corpora.

Reviews were tagged using the Stanford Log-Linear Part-Of-Speech Tagger. Then, tags were broken down into WordNet PoS Tags: *a* (adjective), *n* (noun), *v* (verb), *r* (adverb). Words whose tags were not part of those categories were filtered out. The list of selected words was then stemmed.

Table 3.2 shows quantitative information of the adjective "bad", where *Freq* is the total number of tokens of a (word,tag) pair in each category and corpus, while *Total* is the total number of word tokens in each category and corpus (Total values are constant for all words but

[3] `http://www.stanford.edu/\begingroup\let\relax\relax\endgroup[Pleaseinsert\PrerenderUnicode{Ëœ}intopreamble]cgpotts/data/wordnetscales/`

Word	Tag	Category	Freq	Total	Corpus
bad	a	1	122232	25395214	IMDB
bad	a	2	40491	11755132	IMDB
bad	a	3	37787	13995838	IMDB
bad	a	4	33070	14963866	IMDB
bad	a	5	39205	20390515	IMDB
bad	a	6	43101	27420036	IMDB
bad	a	7	46696	40192077	IMDB
bad	a	8	42228	48723444	IMDB
bad	a	9	29588	40277743	IMDB
bad	a	10	51778	73948447	IMDB
bad	a	1	2100	3419923	Goodreads
bad	a	2	1956	3912625	Goodreads
bad	a	3	2780	6011388	Goodreads
bad	a	4	2298	10187257	Goodreads
bad	a	5	2119	16202230	Goodreads
bad	a	1	1127	699695	OpenTable
bad	a	2	2595	2507147	OpenTable
bad	a	3	2859	4207700	OpenTable
bad	a	4	2544	7789649	OpenTable
bad	a	5	1905	8266564	OpenTable
bad	a	1	1241	3419923	Amazon/Tripadvisor
bad	a	2	791	3912625	Amazon/Tripadvisor
bad	a	3	870	6011388	Amazon/Tripadvisor
bad	a	4	1301	10187257	Amazon/Tripadvisor
bad	a	5	2025	16202230	Amazon/Tripadvisor

Table 3.2: A sample of the collection format for the word ("bad", *a*) in each category

repeated for each one in order to make processing easier). Then, we compute $AvVN$, $AvNVN$, $AvVP$ and $AvNVP$ for each word and obtain the weights ($D_{neg}(w)$ and $D_{pos}(w)$ values) to build the corresponding lexicons for each corpus. Finally, we compute the average of all weights for the same w in order to obtain two cross-domain final lexicons (VERY-NEG and VERY-POS). VERY-NEG contains a list of the most negative words (VN) and a list of words that are not classified as very negative (NVN). In the same way, VERY-POS contains two lists: the most positive words (VP) and the other words that are not very positive (NVP). Both lexicons are freely available. [4]

[4] `https://github.com/almatarneh/LEXICONS`

Through preliminary experiments, we found that the best results were obtained by filtering out words with very low weight ($D <= 0.00000001$), which are values close to zero. This means that we filtered out neutral words, i.e. words without polarity.

In order to ensure that all cases are tested, we created lexicons at two different borderline (B) values: B=1 and B=2. The former is used to determine extreme values in scales from 1 to 5. More precisely, when used B=1, we mean that 1 (very negative) and 5 (very positive) are the extreme scores. The latter parametrization (B=2) is used to define extreme values in scales from 1 to 10: in this case, 1 and 2 are extreme values for the negative scale, while 9 and 10 represent the class of most positive opinions. Each of our two lexicons, VERY-NEG and VERY-POS, consists of two lists derived from different values of B, as shown in Tables 6.1 and 6.2.

	Number of words			VN			NVN		
Lexicon	ADJ	ADV	Total	ADJ	ADV	Total	ADJ	ADV	Total
VERY-NEG B=1	11670	2790	14460	4178	1092	5270	7492	1698	9190
VERY-NEG B=2	11557	2771	14328	4966	1266	6232	6591	1505	8096

Table 3.3: Negative lexicons: total number of words (adjectives and adverbs) for each lexicon, and number of words for each class (VN and NVN)

	Number of words			VP			NVP		
Lexicon	ADJ	ADV	Total	ADJ	ADV	Total	ADJ	ADV	Total
VERY-POS B=1	11402	2769	14171	4721	1163	5884	6681	1606	8287
VERY-POS B=2	11472	2772	14244	5753	1339	7092	5719	1433	7152

Table 3.4: Positive lexicons: total number of words (adjectives and adverbs) for each lexicon, and number of words for each class (VP and NVP) in each lexicon

CHAPTER 4

LINGUISTIC FEATURES AND ITS REPRESENTATION

In this chapter, we will describe the most important linguistic features that we will use in our experiments, which will be presented in detail in Chapter 5 of this thesis.

We have focused on the selection of influential linguistic features taking into account the importance of the quality of the selection of features as a key factor in increasing the efficiency of the classifier in determining the target. N-grams, word embedding, and set of textual features (SOTF) are the linguistic features that we will use in combination with sentiment lexicons we have built using our proposed method presented in Chapter 3.

4.1 N-grams Features

We deal with n-grams based on the occurrence of unigrams and bigrams of words in the document. Unigrams (1g) and bigrams (2g) are valuable to detect specific domain-dependent (opinionated) expressions. The influence of this type of content features has been confirmed by several opinion mining studies (Pang et al., 2002; Zhang et al., 2011; Gerani et al., 2009).

Tripathy et al. (2016) proposed an approach to find the polarity of reviews by converting text into numeric matrices using countvectorizer and TF-IDF, and then using it as input in machine learning algorithms for classification. Martín-Valdivia et al. (2013) combined supervised and unsupervised approaches to get meta-classifier. Frequency-Inverse Document Frequency (TF-IDF), Term Frequency (TF), Term Occurrence (TO), and Binary Occurrence

(BO) were considered as feature representation schemes. SVM outperformed NB for both corpora. TF-IDF was reported as better representation scheme. SVM using TF-IDF without stopword and stemmer yielded the best precision. Paltoglou and Thelwall (2010) examined different unigram weighting schemes and found that some variants of TF/IDF are well suited for Sentiment Analysis.

We assign a weight to all terms by using two representations: TF-IDF and CountVectorizer.

TF-IDF is computed in Equation 4.1.

$$tf/idf_{t,d} = (1 + log(tf_{t,d})) \times log(\frac{N}{df_t}). \tag{4.1}$$

where $tf_{t,d}$ is the term frequency of the term t in the document d, N is the number of documents in the collection and, df_t is the number of documents in the collection containing t.

CountVectorizer transforms the document to token count matrix. First, it tokenizes the document and according to a number of occurrences of each token, a sparse matrix is created. In order to create the Matrix, all stopwords are removed from the document collection. Then, the vocabulary is cleaned up by removing those terms appearing in less than 4 documents to filter out those terms that are too infrequent. To convert the reviews to a matrix of TF-IDF features and to a matrix of token occurrences, we used *sklearn* feature extraction python library.[1] [2]

4.2 Word Embedding

Many deep learning models in NLP need word embedding results as input features. Word embedding is a technique for language modeling and feature learning, which converts words in a vocabulary into vectors of continuous real numbers representing their semantic distribution. The technique commonly involves embedding from a high-dimensional sparse vector space to a lower-dimensional dense vector space. Each dimension of the embedding vector represents a latent feature of a word. The vectors may encode linguistic regularities and patterns of the word contexts. The learning of word embeddings can be done using neural networks.

[1] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[2] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfVectorizer

We used the *doc2vec* algorithm introduced in Le and Mikolov (2014) to represent the reviews. This neural-based representation has been shown to be efficient when dealing with high-dimensional and sparse data (Le and Mikolov, 2014; Dai et al., 2015). Doc2vec learns features from the corpus in an unsupervised manner and provides a fixed-length feature vector as output. Then, the output is fed into a machine-learning classifier. We used a freely available implementation of the doc2vec algorithm included in gensim,[3] which is a free Python library. The implementation of the doc2vec algorithm requires the number of features to be returned (length of the vector). So, we performed a grid search over the fixed vector length 100 (Collobert et al., 2011; Mikolov et al., 2013a,b).

4.3 Set of Textual Features (SOTF)

Many textual features may be used as evidences to detect extreme views: both very positive or very negative alike. In this study, we have extracted some of them to examine to what extent they influence the identification of extreme views. Uppercase characters may indicate that the writer is very upset or affected, so we counted the number of words written in uppercase letters. Also, intensifier words could be a reliable indicator of the existence of extreme views. So, we considered words such as *mostly, hardly, almost, fairly, really, completely, definitely, absolutely, highly, awfully, extremely, amazingly, fully*, and so on.

Furthermore, we took into account negation words such as *no, not, none, nobody, nothing, neither, nowhere, never*, etc. In addition, we also considered elongated words and repeated punctuation such as *sooooo, baaaaad, woooow, goooood, ???, !!!!*, etc.. These textual features have been shown to be effective in many studies related to polarity classification such as Taboada et al. (2011); Kennedy and Inkpen (2006).

4.4 Sentiment Lexicon Features

Sentiment words also called opinion words are considered the primary building block in sentiment analysis as it is an essential resource for most sentiment analysis algorithms, and the first indicator to express positive or negative opinions. In Chapter 3, we described a strategy to build sentiment lexicons from corpora.

[3] `https://radimrehurek.com/gensim/`

Features	Descriptions
N-grams	Unigram TF-IDF(1g) Unigram CountVectorizer(1g) Unigram and Bigram TF-IDF (1g 2g) Unigram and Bigram CountVectorizer (1g 2g)
Doc2Vec (100 Feat.)	Generate vectors for the document
SOTF (8 Feat.)	Number and proportion of negation words in the document Number and proportion of uppercase words in the document Number and proportion of elongated words and punctuation in the document Number and proportion of intensifiers words in the document
VERY-NEG B=1 (4 feat.)	Number and proportion of VN terms in the documents Number and proportion of NVN terms in the documents
VERY-NEG B=2 (4 feat.)	Number and proportion of VN terms in the documents Number and proportion of NVN terms in the documents

Table 4.1: Description of all the considered linguistic features in order to identify the most negative opinions (VN vs. NVN)

Features	Descriptions
N-grams	Unigram TF-IDF(1g) Unigram CountVectorizer(1g) Unigram and Bigram TF-IDF (1g 2g) Unigram and Bigram CountVectorizer (1g 2g)
Doc2Vec (100 Feat.)	Generate vectors for the document
SOTF (8 Feat.)	Number and proportion of negation words in the document Number and proportion of uppercase words in the document Number and proportion of elongated words and punctuation in the document Number and proportion of intensifiers words in the document
VERY-POS B=1 (4 feat.)	Number and proportion of VP terms in the documents Number and proportion of NVP terms in the documents
VERY-POS B=2 (4 feat.)	Number and proportion of VP terms in the documents Number and proportion of NVP terms in the documents

Table 4.2: Description of all the considered linguistic features in order to identify the most positive opinions (VP Vs. NVP)

Tables 4.1 and 4.2 summarizes all the features introduced above with a brief description for each one.

CHAPTER 5

SUPERVISED CLASSIFICATION METHODS BASED ON SENTIMENT LEXICONS

In this chapter, we will examine the efficiency of the automatic construction of a sentiment lexicons that have been built in Chapter 3, Similarly, we will measure the effectiveness of the linguistic features that we explained in Chapter 4.

First, our specific domain lexicon (SPLM) will be compared with two other lexicons to evaluate its performance in the standard task of sentiment classification (positive vs. negative).

Second, VERY-NEG and VERY-POS lexicons will be examined. For this purpose, two types of experiments will be conducted: lexicon comparison and combination of empirical features.

Lexicon comparison: the first experiments with VERY-NEG and VERY-POS aims at performing an indirect evaluation procedure. The indirect evaluation consists of measuring the performance of supervised machine learning classifiers based on the lexicons. Also, we will report an extensive set of experiments aimed to compare our automatic constructed lexicons with other four well-known handcraft lexicons for three binary classification tasks:

- positive vs. negative.
- very negative (VN) *vs.* not very negative opinions (NVN).
- very positive (VP) *vs.* not very positive opinions (NVP).

These tasks can be performed by using classifiers modeled with training data in a supervised strategy. Some linguistic characteristics of documents will be encoded as features in vector representation. These vectors and the corresponding labels feed the classifiers. In order to cover several domains, the experiments were carried out using different datasets.

Combining empirical features : the objective of the second experiment with VERY-NEG and VERY-POS is to investigate the effectiveness and limitations of different linguistic features described in Chapter 4 to identify extreme opinions in the hotels' reviews. Our main contribution is to report an extensive set of experiments aimed to evaluate the relative effectiveness of different linguistic features for two binary classification tasks:

- very negative (VN) *vs.* not very negative opinions (NVN).
- very positive (VP) *vs.* not very positive opinions (NVP).

Figure 5.1 synthesizes and categorizes all the experiments introduced above.

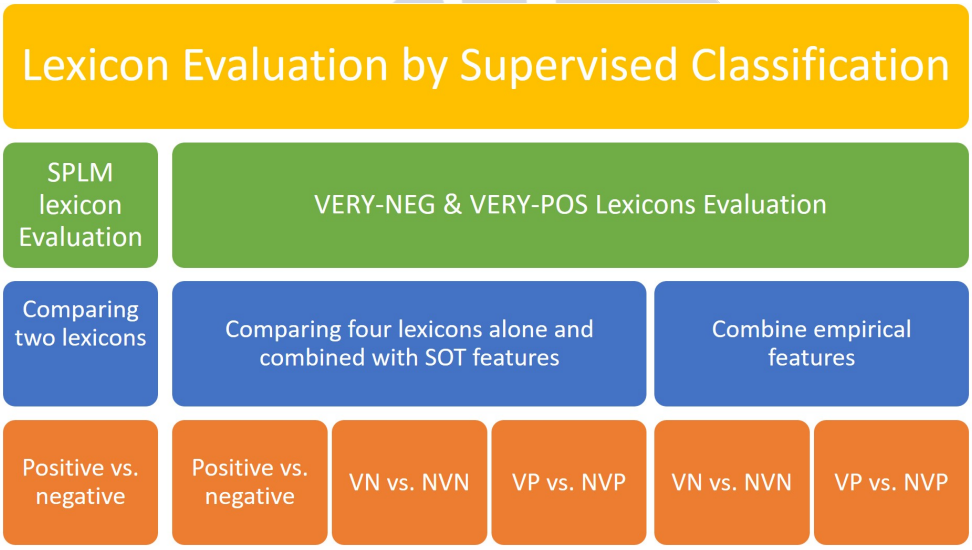

Figure 5.1: The experiments performed to evaluate the performance of lexicons and other features using supervised machine learning classification (SVM).

In the rest of the section, we describe the training method, the evaluation protocol, the test datasets used, as well as the results obtained in the three types of experiments carried out.

5.1 Training and Test

Since we are facing a text classification problem, any existing supervised learning method can be applied. Support Vector Machines (SVM) has been shown to be highly effective at traditional text categorization (Pang et al., 2002). We decided to utilize *scikit*[1] which is an open source machine learning library for Python programming language Pedregosa et al. (2011). We chose SVM as our classifier for all experiments, hence, in this study we will only summarize and discuss results for this learning model. More specifically, we utilized the sklearn.svm.LinearSVC module[2]. Supervised classification requires two samples of documents: training and testing. The training sample will be used to learn various characteristics of the documents and the testing sample was used to predict and next verify the efficiency of our classifier in the prediction. The data set was randomly partitioned into training (75 %) and test (25 %). In contrast to the binary classification of positive and negative views, in case of extreme opinions classification in all collections, the two-class categorization is unbalanced: there are much fewer VN and VP reviews than NVN and NVP ones. Therefore, as recommended in Hsu et al. (2003), we examined the performance by giving more importance to the class of extreme values: both VN and VP. We found that performance was sensitive to the SVM weights which modify the relative cost of misclassifying positive and negative samples. In our analysis, we employed 5_fold cross_validation and the effort was put on optimizing F1 which is computed with respect to VN and VP in the first two tasks (which is the target class):

$$F1 = 2 * \frac{P * R}{P + R} \tag{5.1}$$

where P and R are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{5.2}$$

$$R = \frac{TP}{TP + FN} \tag{5.3}$$

Where TP stands for true positive, FP is false positive, and FN is false negative. To optimize F1, we tried out a grid search approach with exponentially growing sequences of the value of the parameter *class_weight*. More precisely, we tested class_weight with different values: $2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, ..., 2^{10}$. After finding the best value of class_weight within that

[1] http://scikit-learn.org/stable/

[2] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

sequence, we conducted a finer grid search on that better district (e.g. if the optimal value of class_weight is 8, then we test all the neighbors in this region: e.g. 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15 and 16). The class_weight was finally set to the value returning the highest F1 across all these experiments

5.2 Test Datasets

5.2.1 Multi-Domain Sentiment Dataset

This dataset[3] was used in Blitzer et al. (2007a). It contains product reviews taken from Amazon.com for 4 types of products (domains): Kitchen, Books, DVDs, and Electronics.

The star ratings of the reviews are from 1 to 5 stars. In our experiments, we adopted the scale with five categories. In this case, the borderline separating the VN values from the rest was set to 1, which stands for the very negative reviews. The documents in the other four categories were put in the NVN class. According to this borderline value, the VP class was made up of those reviews scored with 5, while the NVP class was built with the rest of reviews.

5.2.2 Sentiment polarity datasets

Sentiment polarity datasets (SPD)[4] consists of 1000 positive and 1000 negative processed reviews. All reviews in this dataset have been extracted from IMDB and Introduced in Pang and Lee (2004b).

5.2.3 Large Movie Review Dataset

Large Movie Review Dataset (LMRD)[5], which was reported in Maas et al. (2011), consists of 50,000 reviews from IMDB, containing over 30 reviews per movie.

The dataset consists of two balanced training and test sets, with 25,000 reviews each. The rating scale is larger than in the previous dataset: it goes from 1 to 10. In this case, the borderline separating the positive values from the negative was set to 4. Concerning the extreme values, if we aim at dividing VN class from the rest, the borderline variable is set to

[3] https://www.cs.jhu.edu/~mdredze/datasets/sentiment/domain_sentiment_data.tar.gz

[4] https://www.cs.cornell.edu/people/pabo/movie-review-data/

[5] http://ai.stanford.edu/~amaas/data/sentiment/

2; so VN reviews were assigned values between 1 and 2. The reviews in the other 8 categories were assigned to the class NVN. The same procedure was carried out on the VP scale.

5.2.4 Hotels Dataset

We obtained the holte dataset from Expedia crowd-sourced data. The HotelExpedia dataset[6] originally contains 6,030 hotels and 381,941 reviews from 11 different hotel locations. The datasets are cleaned and prepared for analysis by applying the following three preprocessing steps: (1) data deduplication operation is performed in order to remove such duplicate reviews; (2) 3-stars reviews were deleted since they tend to contain neutral views; (3) all reviews containing less than three words and blank reviews were also removed. After the above three data cleansing operations, the final datasets consists of 20,000 reviews, being 5,000 for each category: 1, 2, 4 and 5 stars.

Table 5.1 describes the five datasets that were used to evaluate the performance of the lexicons in the sentiment classification task.

Datasets	# of Reviews	Positive	Negative	VN	NVN	VP	NVP
Books	2000	1000	1000	522	1478	731	1269
DVDs	2000	1000	1000	530	1470	714	1286
Electronics	2000	1000	1000	666	1334	680	1320
Kitchens	2000	1000	1000	687	1313	754	1246
LMRD	50000	25000	25000	14708	35292	14338	35662
SPD	2000	1000	1000	-	-	-	-
Hotels	20000	10000	10000	5000	5000	5000	5000

Table 5.1: Size of the five test datasets and the total number of reviews in each class (VN *vs.* NVN) and (VP *vs.* NVP)

5.3 SPLM Lexicons Evaluation

The first experiment aims at comparing our standard polarity lexicon, SPLM, with other two existing lexical resources in the standard task of sentiment analysis. For this purpose, we train a sentiment classifier by making use of simple lexicon-based features, namely: the number of positive and negative terms in the document, and the proportion of positive and negative

[6] `http://ave.dee.isep.ipp.pt/\begingroup\let\relax\relax\endgroup[Pleaseinsert\PrerenderUnicode{Ëœ}intopreamble]1080560/ExpediaDataSet.7z`

terms. We use just lexicon-based features because the purpose of the evaluation is to measure the quality of the given lexicon.

In order to evaluate the performance of the proposed lexicons in a sentiment classification task, we used two datasets, Sentiment polarity datasets (SPD) and Large Movie Review Dataset (LMRD), which we described in Section 5.2.

Lexicon	**Dataset**	**Negative**			**Positive**		
		P	*R*	*F*1	*P*	*R*	*F*1
SPLM	SPD	0.84	0.81	**0.83**	0.81	0.84	**0.83**
	LMRD	0.77	0.75	**0.76**	0.75	0.77	**0.76**
SO-CAL	SPD	0.69	0.69	0.69	0.67	0.67	0.67
	LMRD	0.72	0.68	0.70	0.69	0.73	0.71
SentiWords	SPD	0.72	0.69	0.70	0.69	0.72	0.71
	LMRD	0.69	0.65	0.67	0.67	0.67	0.70

Table 5.2: Results in terms of precision (P), recall (R), and F_1 scores for Positive and Negative classification. The best F_1 in each datset is highlighted (in bold)

The experimental results are shown in Table 5.2. By comparing the F1-score obtained by the three lexicons, we may conclude that the lexicon we automatically generated, SPLM, consistently outperforms the other manual lexicons on the two datasets.

It is worth noticing that SO-CAL and SentiWords are general-purpose polarity lexicons, while SPLM is a domain-specific resource. This might explain why our lexicon performs better. However, we should point out that SPLM is the result of an automatic method while the other resources were made manually.

5.4 VERY-NEG and VERY-POS Lexicon Evaluation

In the following experiments, we evaluate the efficiency of our extreme lexicons: VERY-NEG and VERY-POS.

5.4.1 Comparison of Lexicons

In order to cover several domains, the experiments were carried out using the Multi-Domain Sentiment Dataset defined in Section 5.2. Our lexicons (VERY-NEG and VERY-POS) were evaluated and compared with other existing lexicons in the two tasks of classifying reviews. As mentioned earlier, there are many popular and available sentiment lexicons. We used two types of them: First, lexicons assigning PoS tags to lemmas, such as SO-CAL and SentiWords. In our experiments, only adjectives and adverbs were compared. Second, lexicons without POS tags: Hu & Liu Opinion Lexicon and AFINN-111. Six lexicons will be compared depending on each task: the two lexicons we automatically built using our strategy, called VERY-NEG and VERY-POS, and four manual resources: SO-CAL (Taboada et al., 2011), SentiWords (Gatti et al., 2016), Hu & Liu Opinion Lexicon (Hu and Liu, 2004; Liu et al., 2005), and AFINN-111 (Nielsen, 2011).

5.4.1.1 Positive *vs.* Negative

Tables 5.3, 5.4, 5.5, and 5.6 summarize the polarity classification results, in terms of P, R, and F1, for all dataset collections (Book, DVD, Electronic, and Kitchen), by making use of all compared lexicons (including our VERY-NEG and VERY-POS), within the framework of a standard positive vs negative classification task.

Lexicon	BOOK						
	Negative			Positive			AVG F1
	P	R	F1	P	R	F1	
VERY-NEG B=1	0.75	0.70	0.72	0.71	0.76	**0.74**	**0.73**
VERY-NEG B=2	0.75	0.71	0.73	0.71	0.75	0.73	0.73
VERY-POS B=1	0.71	0.69	0.70	0.69	0.72	0.70	0.70
VERY-POS B=2	0.76	0.77	**0.76**	0.76	0.75	**0.75**	**0.76**
SO-CAL	0.72	0.60	0.66	0.65	0.76	0.70	0.68
SentiWorrds	0.70	0.61	0.65	0.64	0.73	0.68	0.67
Opinion Lexicon	0.68	0.67	0.68	0.67	0.67	0.67	0.68
AFINN-111	0.69	0.65	0.67	0.66	0.70	0.68	0.68
VERY-NEG B=1 +SOTF	0.75	0.71	0.73	0.71	0.75	0.73	0.73
VERY-NEG B=2 +SOTF	0.76	0.74	0.75	0.74	0.75	0.74	0.75
VERY-POS B=1 +SOTF	0.74	0.72	0.73	0.72	0.73	0.73	0.73
VERY-POS B=2 +SOTF	0.76	0.76	0.76	0.75	0.75	**0.75**	**0.76**
SO-CAL +SOTF	0.76	0.63	0.69	0.68	0.79	0.73	0.71
SentiWorrds +SOTF	0.72	0.63	0.67	0.66	0.74	0.70	0.69
Opinion Lexicon +SOTF	0.73	0.73	0.73	0.72	0.73	0.72	0.73
AFINN-111 +SOTF	0.73	0.71	0.72	0.71	0.72	0.71	0.72

Table 5.3: Polarity classification results for Book collection with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R), F1 scores of all F1 for *negative* and *positive* classes. The best F1 in each lexicon is highlighted (in bold).

Lexicon	DVD						
	Negative			Positive			AVG F1
	P	R	F1	P	R	F1	
VERY-NEG B=1	0.67	0.67	0.67	0.66	0.67	0.66	0.67
VERY-NEG B=2	0.70	0.70	0.70	0.69	0.68	0.69	0.70
VERY-POS B=1	0.70	0.68	0.69	0.68	0.70	0.69	0.69
VERY-POS B=2	0.71	0.70	0.70	0.69	0.70	0.70	0.70
SO-CAL	0.68	0.65	0.67	0.66	0.69	0.67	0.67
SentiWorrds	0.69	0.66	0.67	0.66	0.70	0.68	0.68
Opinion Lexicon	0.75	0.72	**0.74**	0.72	0.75	**0.74**	**0.74**
AFINN-111	0.74	0.69	0.72	0.70	0.75	0.72	0.72
VERY-NEG B=1 +SOTF	0.67	0.63	0.65	0.64	0.68	0.66	0.66
VERY-NEG B=2 +SOTF	0.70	0.71	0.71	0.70	0.69	0.69	0.70
VERY-POS B=1 +SOTF	0.71	0.70	0.70	0.69	0.70	0.70	0.70
VERY-POS B=2 +SOTF	0.71	0.70	0.70	0.69	0.70	0.70	0.70
SO-CAL +SOTF	0.70	0.67	0.68	0.67	0.71	0.69	0.69
SentiWorrds +SOTF	0.70	0.65	0.68	0.67	0.72	0.69	0.69
Opinion Lexicon +SOTF	0.75	0.74	**0.74**	0.73	0.74	**0.74**	**0.74**
AFINN-111 +SOTF	0.74	0.72	0.73	0.72	0.74	0.73	0.73

Table 5.4: Polarity classification results for DVD collection with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R), F1 scores of all F1 for *negative* and *positive* classes. The best F1 in each lexicon is highlighted (in bold).

Lexicon	Electronic						
	Negative			Positive			AVG F1
	P	R	F1	P	R	F1	
VERY-NEG B=1	0.74	0.70	0.72	0.71	0.74	0.72	0.72
VERY-NEG B=2	0.72	0.74	0.73	0.72	0.70	0.71	0.72
VERY-POS B=1	0.65	0.69	0.67	0.66	0.62	0.64	0.66
VERY-POS B=2	0.69	0.68	0.69	0.68	0.69	0.68	0.69
SO-CAL	0.70	0.68	0.69	0.68	0.70	0.69	0.69
SentiWorrds	0.72	0.67	0.69	0.68	0.73	0.70	0.70
Opinion Lexicon	0.75	0.75	**0.75**	0.74	0.74	**0.74**	**0.75**
AFINN-111	0.75	0.72	0.73	0.72	0.75	0.73	0.73
VERY-NEG B=1 +SOTF	0.73	0.71	0.72	0.71	0.72	0.72	0.72
VERY-NEG B=2 +SOTF	0.75	0.73	0.74	0.73	0.75	0.74	0.74
VERY-POS B=1 +SOTF	0.70	0.69	0.70	0.69	0.70	0.69	0.70
VERY-POS B=2 +SOTF	0.71	0.69	0.70	0.69	0.71	0.70	0.70
SO-CAL +SOTF	0.70	0.68	0.69	0.68	0.70	0.69	0.69
SentiWorrds +SOTF	0.72	0.67	0.69	0.68	0.73	0.70	0.70
Opinion Lexicon +SOTF	0.74	0.76	0.75	0.74	0.73	0.74	0.75
AFINN-111 +SOTF	0.76	0.76	**0.76**	0.75	0.75	**0.75**	**0.76**

Table 5.5: Polarity classification results for Electronic collection with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R), F1 scores of all F1 for *negative* and *positive* classes. The best F1 in each lexicon is highlighted (in bold).

Lexicon	Kitchen						
	Negative			Positive			AVG F1
	P	R	F1	P	R	F1	
VERY-NEG B=1	0.70	0.72	**0.71**	0.70	0.67	0.69	0.70
VERY-NEG B=2	0.70	0.71	**0.71**	0.70	0.68	0.69	0.70
VERY-POS B=1	0.70	0.67	0.68	0.67	0.70	0.69	0.69
VERY-POS B=2	0.71	0.70	**0.71**	0.69	0.71	**0.70**	**0.71**
SO-CAL	0.67	0.71	0.69	0.68	0.63	0.66	0.68
SentiWorrds	0.69	0.68	0.69	0.67	0.68	0.68	0.69
Opinion Lexicon	0.71	0.69	0.70	0.69	0.72	**0.70**	0.70
AFINN-111	0.71	0.69	0.70	0.69	0.70	0.69	0.70
VERY-NEG B=1 +SOTF	0.69	0.72	0.70	0.69	0.67	0.68	0.69
VERY-NEG B=2 +SOTF	0.71	0.74	0.72	0.71	0.68	0.70	0.71
VERY-POS B=1 +SOTF	0.70	0.70	0.70	0.69	0.69	0.69	0.70
VERY-POS B=2 +SOTF	0.70	0.74	0.72	0.72	0.68	0.70	0.71
SO-CAL +SOTF	0.70	0.72	0.71	0.70	0.68	0.69	0.70
SentiWorrds +SOTF	0.70	0.69	0.69	0.68	0.70	0.69	0.69
Opinion Lexicon +SOTF	0.75	0.72	**0.73**	0.72	0.75	**0.73**	**0.73**
AFINN-111 +SOTF	0.71	0.72	0.72	0.71	0.70	0.70	0.71

Table 5.6: Polarity classification results for Kitchen collection with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R), F1 scores of all F1 for *negative* and *positive* classes. The best F1 in each lexicon is highlighted (in bold).

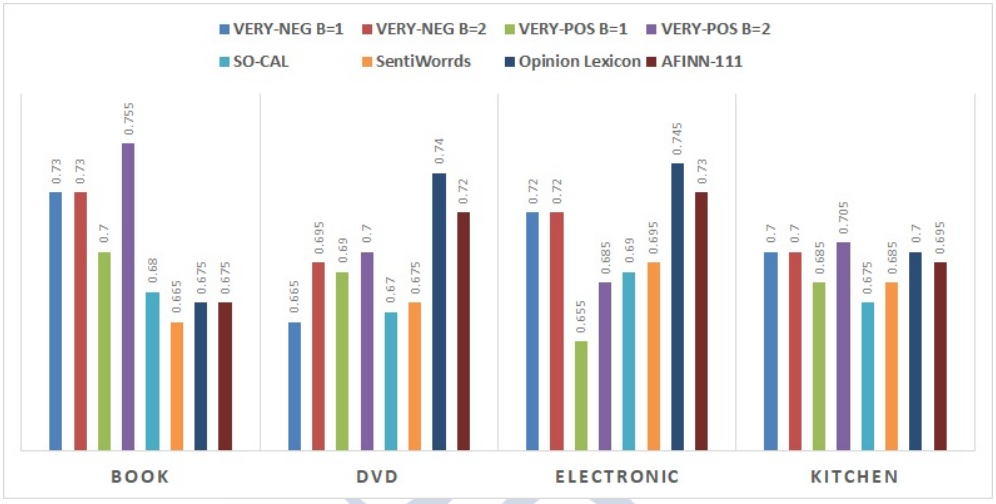

Figure 5.2: Polarity classification results for all collections with all lexicons alone in terms of average of F1 scores for *negative* and *positive* classes.

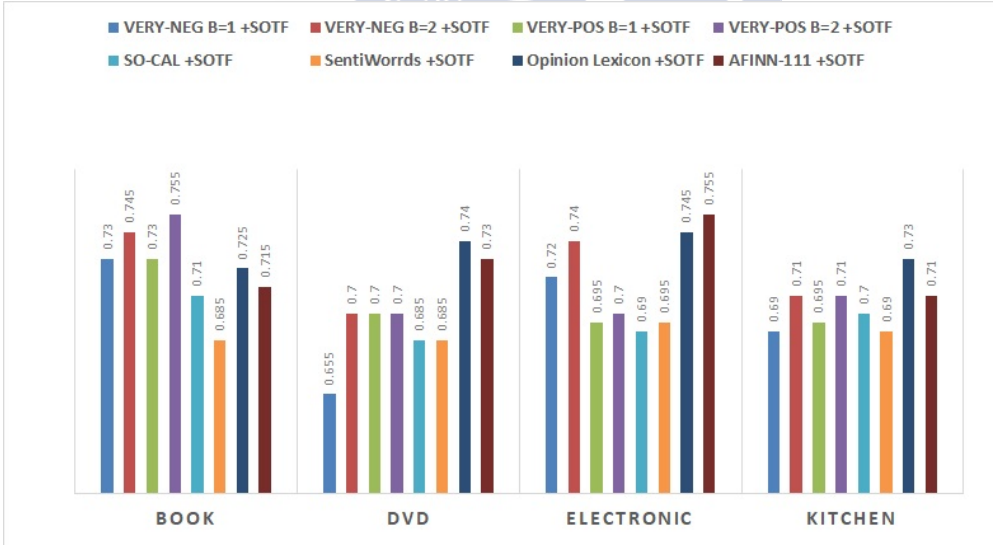

Figure 5.3: Polarity classification results for all collections with all lexicons with SOTF in terms of average of F1 scores for *negative* and *positive* classes.

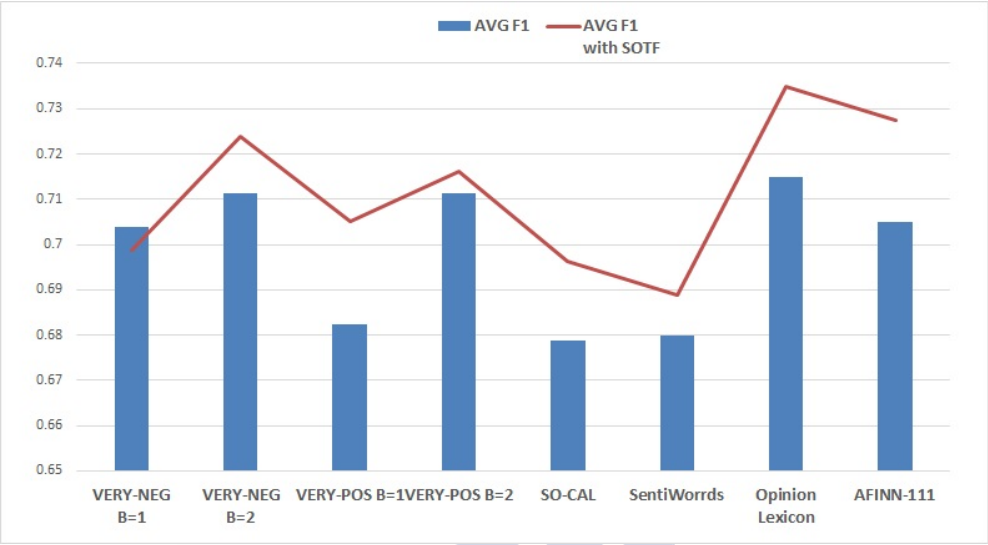

Figure 5.4: Comparison between the polarity classification results for all collections with all lexicons alone and with SOTF, regarding the average of all F1 for *positive* and *negative* classes.

The results represented in tables 5.3, 5.4, 5.5, and 5.6 show the following trends: The performance of all lexicons in all datasets is almost alike. Considering the average of the four datasets (last column in tables), our automatic lexicon outperformed the other lexicons in both book and kitchen datasets while the performance of the lexicon Opinion Lexicon was better in the rest of the collections: DVD and electronic.

Figs. 5.2, 5.3 and 5.4 depict more clearly that polarity classification results regarding the average of all F1 for all collections are improved when the lexicons are combined with linguistic features (SOTF). This improvement happens for all lexicons, which shows that such features help to boost the overall sentiment classification.

5.4.1.2 Very Negative Classification (VN vs NVN)

Tables 5.7, 5.8, 5.9, and 5.10 show the results of polarity classification for all datasets Book, DVD, Electronic and Kitchen by take advantage of all examined lexicons in two forms: lexical features alone and integrated with SOTF in terms of Precision (P), Recall (R) and F1 scores for very negative class (VN).

	BOOK		
Lexicon	P	R	F1
VERY-NEG B=1	0.46	0.76	0.58
VERY-NEG B=2	0.48	0.80	**0.60**
SO-CAL	0.44	0.64	0.52
SentiWords	0.41	0.66	0.51
Opinion Lexicon	0.42	0.66	0.52
AFINN-111	0.44	0.66	0.52
VERY-NEG B=1 +SOTF	0.48	0.75	0.59
VERY-NEG B=2 +SOTF	0.50	0.81	**0.62**
SO-CAL +SOTF	0.47	0.68	0.55
SentiWords +SOTF	0.44	0.66	0.53
Opinion Lexicon +SOTF	0.47	0.74	0.58
AFINN-111 +SOTF	0.47	0.69	0.56

Table 5.7: Polarity classification results for Book dataset with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R) and F1 scores for *very negative* class (VN). The best F1 is highlighted (in bold).

	DVD		
Lexicon	P	R	F1
VERY-NEG B=1	0.57	0.64	**0.60**
VERY-NEG B=2	0.46	0.78	0.58
SO-CAL	0.45	0.73	0.56
SentiWords	0.42	0.66	0.52
Opinion Lexicon	0.48	0.80	**0.60**
AFINN-111	0.49	0.78	**0.60**
VERY-NEG B=1 +SOTF	0.57	0.62	0.60
VERY-NEG B=2 +SOTF	0.46	0.76	0.58
SO-CAL +SOTF	0.47	0.75	0.58
SentiWords +SOTF	0.44	0.68	0.53
Opinion Lexicon +SOTF	0.59	0.64	**0.61**
AFINN-111 +SOTF	0.61	0.61	**0.61**

Table 5.8: Polarity classification results for DVD dataset with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R) and F1 scores for *very negative* class (VN). The best F1 is highlighted (in bold).

	Electronic		
Lexicon	P	R	F1
VERY-NEG B=1	0.52	0.86	**0.65**
VERY-NEG B=2	0.49	0.87	0.63
SO-CAL	0.55	0.71	0.62
SentiWords	0.54	0.67	0.60
Opinion Lexicon	0.5	0.85	0.63
AFINN-111	0.48	0.87	0.62
VERY-NEG B=1 +SOTF	0.53	0.86	**0.66**
VERY-NEG B=2 +SOTF	0.52	0.86	0.64
SO-CAL +SOTF	0.49	0.85	0.62
SentiWords +SOTF	0.48	0.84	0.61
Opinion Lexicon +SOTF	0.52	0.85	0.64
AFINN-111 +SOTF	0.51	0.9	0.65

Table 5.9: Polarity classification results for Electronic datasetwith all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R) and F1 scores for *very negative* class (VN). The best F1 is highlighted (in bold).

	Kitchen		
Lexicon	P	R	F1
VERY-NEG B=1	0.53	0.72	0.61
VERY-NEG B=2	0.54	0.72	**0.62**
SO-CAL	0.43	0.92	0.58
SentiWords	0.45	0.93	0.61
Opinion Lexicon	0.44	0.94	0.60
AFINN-111	0.43	0.94	0.59
VERY-NEG B=1 +SOTF	0.53	0.72	0.61
VERY-NEG B=2 +SOTF	0.55	0.75	**0.64**
SO-CAL +SOTF	0.44	0.93	0.60
SentiWords +SOTF	0.45	0.93	0.61
Opinion Lexicon +SOTF	0.44	0.93	0.60
AFINN-111 +SOTF	0.45	0.92	0.61

Table 5.10: Polarity classification results for Kitchen dataset with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R) and F1 scores for *very negative* class (VN). The best F1 is highlighted (in bold).

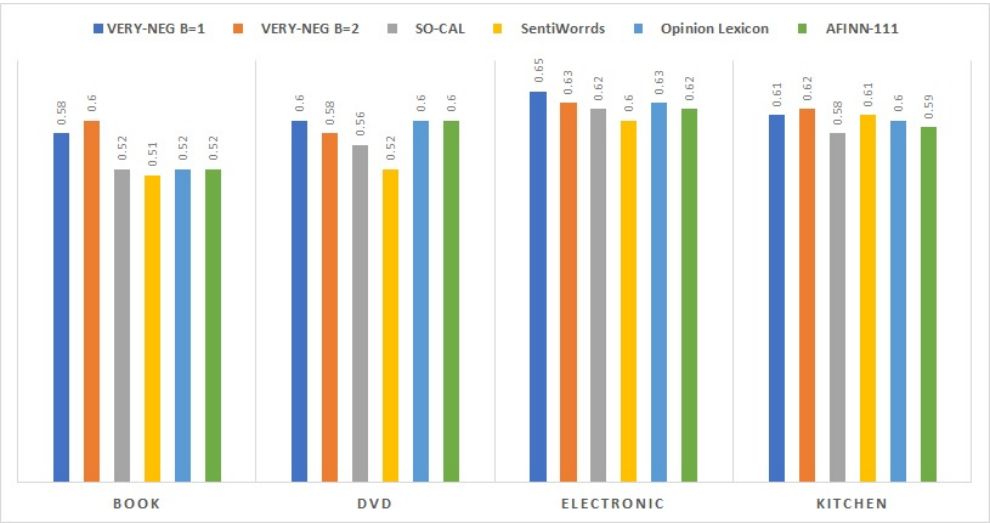

Figure 5.5: Polarity classification results for all collections with all lexicons in terms of F1 scores for *very negative* class (VN).

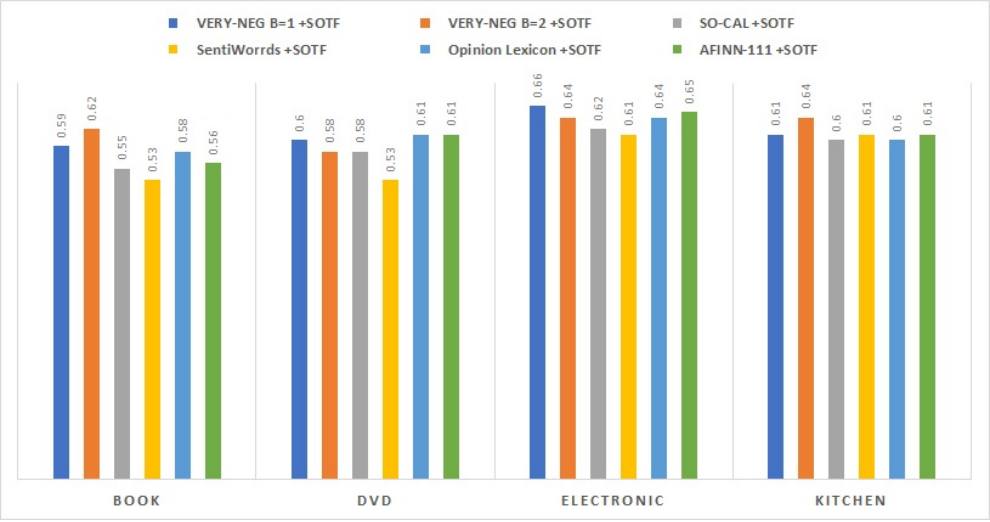

Figure 5.6: Polarity classification results for all collections with all lexicons with SOTF in terms of F1 scores for *very negative* class (VN).

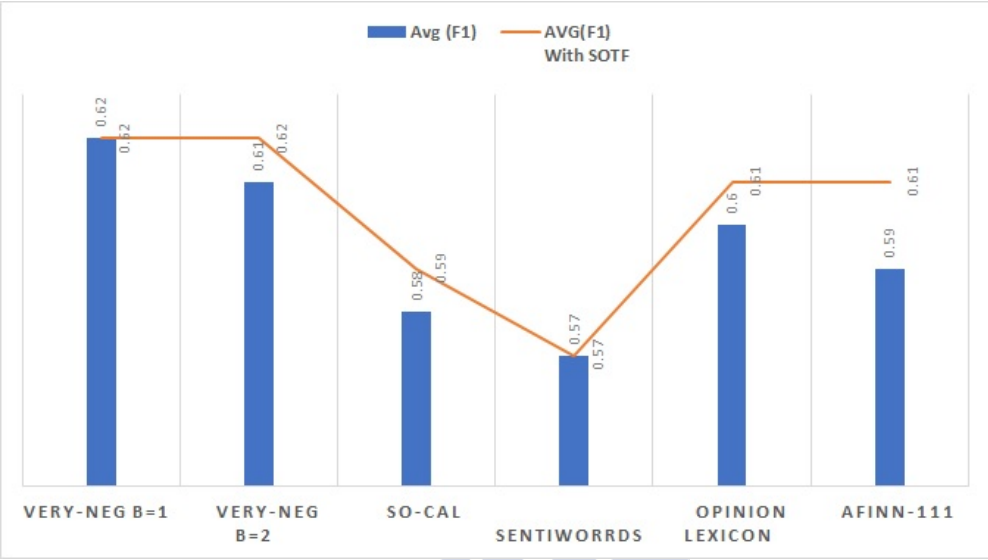

Figure 5.7: Comparison between the polarity classification results for all collections with all lexicons alone and with SOTF, regarding the average of all F1 for *very negative* class (VN).

Considering the average of the four datasets (last column in Tables 5.7, 5.8, 5.9, and 5.10), the classifier configured with our lexicons outperforms the same classifier trained with the manual resources. The same thing happens when we add SOTF features to the classifier as Figs. 5.5, 5.6 and 5.7 shows. However, it is worth noting that in two of the datasets, namely DVD and Electronic, the results seem more mitigated, which is going to require a deeper analysis of errors.

5.4.1.3 Very Positive Classification (VP vs NVP)

Lexicon	BOOK		
	P	R	F1
VERY-POS B=1	0.56	0.80	**0.66**
VERY-POS B=2	0.57	0.78	**0.66**
SO-CAL	0.41	0.94	0.57
SentiWords	0.40	0.94	0.56
Opinion Lexicon	0.41	0.92	0.57
AFINN-111	0.40	0.93	0.56
VERY-POS B=1 +SOTF	0.58	0.80	**0.67**
VERY-POS B=2 +SOTF	0.58	0.77	0.66
SO-CAL +SOTF	0.44	0.93	0.59
SentiWords +SOTF	0.43	0.90	0.58
Opinion Lexicon +SOTF	0.44	0.88	0.59
AFINN-111 +SOTF	0.42	0.89	0.57

Table 5.11: Polarity classification results for Book dataset with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R) and F1 scores for *very positive* class (VP). The best F1 is highlighted (in bold).

Lexicon	DVD		
	P	R	F1
VERY-POS B=1	0.47	0.85	0.60
VERY-POS B=2	0.45	0.78	0.57
SO-CAL	0.43	0.91	0.58
SentiWords	0.42	0.94	0.58
Opinion Lexicon	0.51	0.76	**0.61**
AFINN-111	0.43	0.91	0.58
VERY-POS B=1 +SOTF	0.47	0.83	0.60
VERY-POS B=2 +SOTF	0.45	0.78	0.57
SO-CAL +SOTF	0.44	0.89	0.59
SentiWords +SOTF	0.42	0.90	0.58
Opinion Lexicon +SOTF	0.52	0.75	**0.62**
AFINN-111 +SOTF	0.49	0.76	0.59

Table 5.12: Polarity classification results for DVD dataset with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R) and F1 scores for *very positive* class (VP). The best F1 is highlighted (in bold).

Lexicon	Electronic		
	P	R	F1
VERY-POS B=1	0.51	0.65	0.57
VERY-POS B=2	0.50	0.69	0.58
SO-CAL	0.49	0.69	0.57
SentiWords	0.44	0.87	0.58
Opinion Lexicon	0.45	0.87	**0.60**
AFINN-111	0.42	0.88	0.57
VERY-POS B=1 +SOTF	0.52	0.70	**0.60**
VERY-POS B=2 +SOTF	0.52	0.71	**0.60**
SO-CAL +SOTF	0.44	0.86	0.58
SentiWords +SOTF	0.45	0.87	0.59
Opinion Lexicon +SOTF	0.46	0.85	0.59
AFINN-111 +SOTF	0.43	0.88	0.58

Table 5.13: Polarity classification results for Electronic dataset with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R) and F1 scores for *very positive* class (VP). The best F1 is highlighted (in bold).

Lexicon	Kitchen		
	P	R	F1
VERY-POS B=1	0.52	0.73	0.60
VERY-POS B=2	0.46	0.89	**0.61**
SO-CAL	0.44	0.93	0.59
SentiWords	0.42	0.95	0.58
Opinion Lexicon	0.44	0.95	**0.61**
AFINN-111	0.44	0.92	0.60
VERY-POS B=1 +SOTF	0.52	0.80	**0.63**
VERY-POS B=2 +SOTF	0.52	0.77	0.62
SO-CAL +SOTF	0.47	0.89	0.61
SentiWords +SOTF	0.45	0.85	0.59
Opinion Lexicon +SOTF	0.46	0.91	0.61
AFINN-111 +SOTF	0.48	0.84	0.61

Table 5.14: Polarity classification results for Kitchen dataset with all lexicons alone and combined with SOTF, in terms of Precision (P), Recall (R) and F1 scores for *very positive* class (VP). The best F1 is highlighted (in bold).

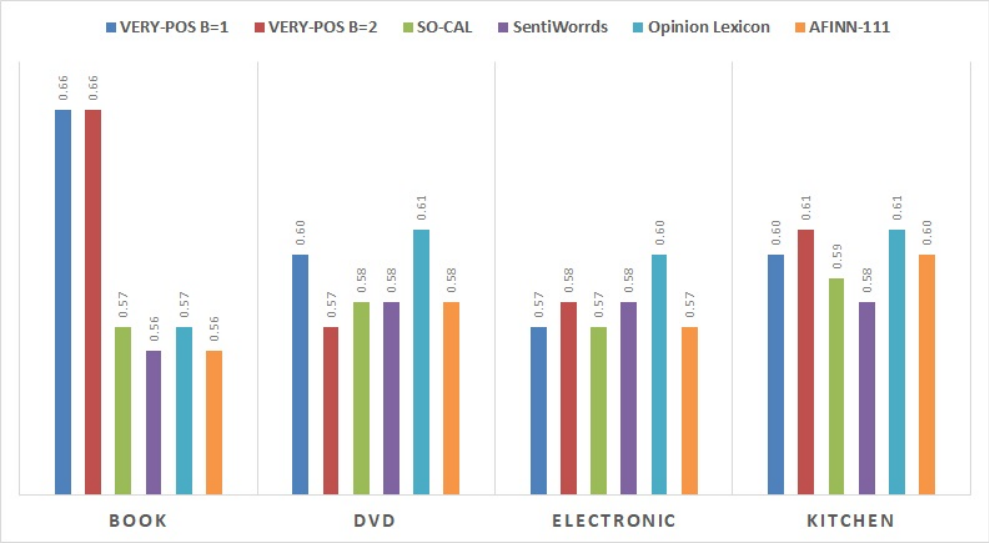

Figure 5.8: Polarity classification results for all collections with all lexicons in terms of average of F1 scores for *very positive* class (VP).

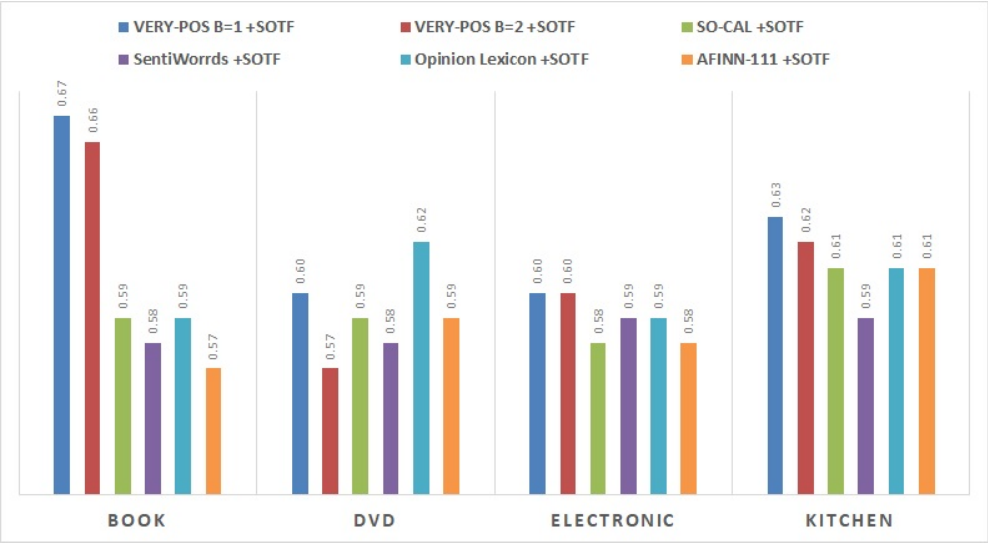

Figure 5.9: Polarity classification results for all collections with all lexicons with SOTF in terms of F1 scores for *very positive* class (VP).

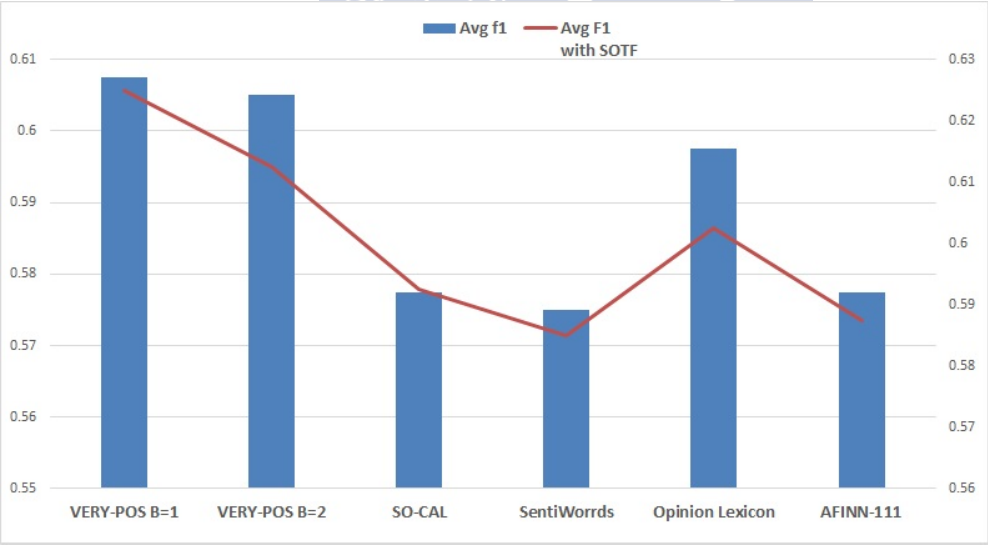

Figure 5.10: Comparison between the polarity classification results for all collections with all lexicons alone and with SOTF, regarding the average of all F1 for *very positive* class (VP)

In the experiments carried out so far, our lexicons have shown high efficiency in tasks related to the identification of extreme views, more precisely, VN vs. NVN and VP vs. NVP tasks. Although our lexicons did not outperform all dictionaries in standard supervised sentiment classification (positive vs. negative), their performance is better than that achieved by most of the other dictionaries we have compared in our experiments with all datasets.

The single most striking observation to emerge from the data comparison was that the combination of SOTF with lexicon features showed an unstable impact to improve the performance of the classifier to identify very positive opinions (see Fig. 5.10), although, it showed significant influence in the two other tasks, namely VN vs. NVN and standard sentiment classification (positive vs. negative). A possible explanation for this might be that the SOTF we have selected are more biased to be evidence of the very negative reviews than being an indication of the most positive opinions. For example, elongated words, negation words, and uppercase words, which take part of our SOFT such as we described in Section 4.3 of Chapter 4, are cues for finding very negative opinions, and not positive ones.

5.4.2 Combination of Empirical Features

This section aimed to evaluate the relative effectiveness of different linguistic features (N-grams, word embedding, polarity lexicons, and set of textual features (SOTF)) which we described in Chapter 4 for two binary classification tasks:

– very negative *vs.* not very negative opinions

– very positive *vs.* not very positive opinions

It is worth noting that we built our lexicons VERY-NEG and VERY-POS for this experiment in in the same way as we built our lexicons in subsection 3.2.1 of Chapter 3. However, in this case, as corpus resource, we only use the hotels and restaurants reviews from OpenTable and Tripadvisor.

5.4.2.1 Very Negative Classification (VN vs NVN)

Table 5.15 shows the performance of very negative classification (VN vs. NVN) performed on our data collection. In these experiments, we combine each n-gram model with the rest of features. The n-gram models are unigrams (1g) and unigrams with bigrams (1g 2g), each one weighted with TF-IDF and CountVector. These models were considered as baselines.

Then, we combined each baseline with one of the rest of features: namely, doc2vec, SOTF, *VERY-NEG* B=1, *VERY-NEG* B=2, (see Table 4.1). Moreover, we also combined all features with each baseline (All).

	VN			**NVN**			
Features	*P*	*R*	*F*1	*P*	*R*	*F*1	s-test
1g(TF-IDF)	0.75	0.64	0.69	0.89	0.93	0.91	
+ Doc2Vec	0.77	0.70	0.73	0.91	0.93	0.92	$\gg$
+ SOTF	0.76	0.66	0.71	0.90	0.93	0.91	$>$
+ *VERY-NEG* B=1	0.76	0.65	0.70	0.89	0.93	0.91	$\sim$
+ *VERY-NEG* B=2	0.76	0.66	0.71	0.89	0.93	0.91	$\sim$
+ All	0.78	0.72	0.75	0.91	0.93	0.92	$\gg$
1g(CountVector)	0.67	0.66	0.66	0.89	0.90	0.89	
+ Doc2Vec	0.72	0.70	0.71	0.91	0.91	0.91	$\gg$
+ SOTF	0.68	0.68	0.68	0.90	0.90	0.90	$>$
+ *VERY-NEG* B=1	0.68	0.67	0.67	0.89	0.90	0.90	$\sim$
+ *VERY-NEG* B=2	0.68	0.66	0.67	0.89	0.90	0.90	$\sim$
+ All	0.74	0.71	0.73	0.91	0.92	0.91	$\gg$
1g 2g(TF-IDF)	0.77	0.62	0.69	0.89	0.94	0.91	
+ Doc2Vec	0.79	0.69	0.74	0.90	0.94	0.92	$\gg$
+ SOTF	0.78	0.64	0.70	0.89	0.94	0.92	$>$
+ *VERY-NEG* B=1	0.68	0.73	0.70	0.89	0.94	0.91	$\sim$
+ *VERY-NEG* B=2	0.79	0.63	0.70	0.89	0.94	0.92	$\sim$
+ All	0.81	0.72	0.76	0.91	0.94	0.93	$\gg$
1g 2g(CountVector)	0.69	0.65	0.67	0.89	0.91	0.90	
+ Doc2Vec	0.75	0.70	0.73	0.91	0.93	0.92	$\gg$
+ SOTF	0.71	0.67	0.69	0.90	0.91	0.90	$\gg$
+ *VERY-NEG* B=1	0.71	0.66	0.68	0.89	0.91	0.90	$\sim$
+ *VERY-NEG* B=2	0.71	0.66	0.68	0.89	0.91	0.90	$\sim$
+ All	0.71	0.68	0.69	0.90	0.91	0.91	$>$

Table 5.15: Polarity classification results, in terms of precision, recall, and F1 scores of VN and NVN. For each n-gram-based model the best performance for each metric is in bold. The symbol "$\gg$" and "$\ll$" indicates a significant improvement with respect to the n-gram-based baselines, with p-value ≤ 0.01. The symbol "$>$" or "$<$" means that the 0.01<p-value$\leq$0.05. "$\sim$" indicates that the difference was not statistically significant (p-value $> .05$).

5.4.2.2 Very Positive Classification (VP vs NVP)

In Table 5.16, we report the performance of very positive classification (VP vs. NVP) on our dataset. As we did with the very negative classification, n-gram-based classifiers were

regarded as baselines, and we examined the association of various combinations of features into the baseline classifiers, including configurations combining all features.

	MP			NMP			
Features	*P*	*R*	*F*1	*P*	*R*	*F*1	s-test
1g(TF-IDF)	0.87	0.83	0.85	0.94	0.96	0.95	
+ Doc2Vec	0.89	0.85	0.87	0.95	0.96	0.96	$>$
+ SOTF	0.88	0.87	0.87	0.95	0.96	0.96	$\gg$
+ *VERY-POS* B=1	0.87	0.83	0.85	0.94	0.96	0.95	$\sim$
+ *VERY-POS* B=2	0.87	0.84	0.86	0.95	0.96	0.95	$\sim$
+ All	0.89	0.87	0.88	0.96	0.96	0.96	$\gg$
1g(CountVector)	0.81	0.79	0.80	0.93	0.93	0.93	
+ Doc2Vec	0.85	0.84	0.85	0.95	0.95	0.95	$\gg$
+ SOTF	0.84	0.83	0.83	0.94	0.94	0.94	$\gg$
+ *VERY-POS* B=1	0.82	0.80	0.81	0.93	0.94	0.94	$>$
+ *VERY-POS* B=2	0.81	0.80	0.81	0.93	0.94	0.93	$\sim$
+ All	0.86	0.84	0.85	0.94	0.95	0.95	$\gg$
1g 2g(TF-IDF)	0.88	0.84	0.86	0.94	0.96	0.95	
+ Doc2Vec	0.89	0.86	0.87	0.95	0.96	0.96	$\sim$
+ SOTF	0.88	0.87	0.88	0.96	0.96	0.96	$\gg$
+ *VERY-POS* B=1	0.88	0.84	0.86	0.95	0.96	0.95	$\sim$
+ *VERY-POS* B=2	0.88	0.84	0.86	0.95	0.96	0.95	$\sim$
+ All	0.91	0.89	0.90	0.96	0.97	0.97	$\gg$
1g 2g(CountVector)	0.83	0.81	0.82	0.93	0.94	0.94	
+ Doc2Vec	0.86	0.85	0.86	0.95	0.95	0.95	$\gg$
+ SOTF	0.86	0.84	0.85	0.94	0.95	0.95	$\gg$
+ *VERY-POS* B=1	0.84	0.81	0.82	0.94	0.94	0.94	$\sim$
+ *VERY-POS* B=2	0.84	0.81	0.83	0.94	0.95	0.94	$\sim$
+ All	0.89	0.88	0.88	0.96	0.96	0.96	$\gg$

Table 5.16: Polarity classification results, in terms of precision, recall, and F1 scores of VP and NVP. For each n-gram-based classifier, the best performance for each metric is bolded. The symbol "$\gg$" or "$\ll$" stands for significant improvement with respect to n-gram-based baselines, with p-value ≤ 0.01. The symbol "$>$" or "$<$" means that the $0.01<$p-value≤ 0.05. "$\sim$" indicates that the difference was not statistically significant (p-value $> .05$).

The results depicted by tables 5.15 and 5.16 show the following trends. Concerning the classification of not very extreme opinions (NVN and NVP), the baseline approaches are already very accurate and, so, the use of the rest of features does not provide any significant improvement. By contrast, the classification of very extreme opinions is a more tough task

in which the baselines are outperformed by some of the other features we have tested. The last column in both tables shows the significant differences concerning only VN and VP classifications So, significant tests are shown for classification of extreme opinions. In the case of non extreme opinions, there are no significant improvements when we combine different features.

To detect extreme opinions (both very negative and very positive), the most valuable features are textual features (SOTF) and embeddings (Doc2Vec). However, Doc2Vec is more beneficial to detect the very negative reviews (Table 5.16), while SOTF performs better with the very positive ones (Table 5.16). Both types of features lead to statistically significant improvements when they are combined with the baselines (n-gram representations). This confirms the valuable information provided by Doc2Vec and SOTF to detect the most extreme reviews. Lexicon-based features slightly improve the baselines but not in a significant way.

Besides, in all cases the combination of all features always yield significant improvements with regard to the baselines. Finally, it is worth noting that none of the features hurts the overall performance.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CHAPTER 6

UNSUPERVISED CLASSIFICATION METHODS BASED ON SENTIMENT LEXICON

6.1 Sentiment classification

Sentiment analysis typically works at three levels of granularity, namely, document level, sentence level, and aspect level. We are involved with document-level classification and two polarity classes: extreme *vs.* non-extreme opinions. Our unsupervised sentiment classification is carried out as follows. First, a part-of-speech tagger is applied to extract adjectives and adverbs from reviews. Then, the algorithm plotted in Figs 6.1 and 6.3 is applied. This is a basic word-matching scheme to carried out unsupervised sentiment classification. In particular, the sentiment polarity of a word is obtained from the sentiment lexicon built in the previous step. In the case of classification between VN and NVN, the algorithm in Fig 6.1 assigns -1 to VN words and +1 to NVN. On the other hand, in the case of classification between the VP and NVP, the algorithm assigns +1 to VP words and -1 to NVP as in Fig 6.3.

The overall sentiment score of a document is simply computed as the sum of the sentiment scores of the words in the document.

In order to cover several domains, the experiments were carried out using different datasets, including books, DVD, electronics, housewares, and movie reviews. In our experiments, we automatically built two polarity lexicons using the strategy defined above in Chapter3. Our

lexicons were evaluated and compared with other existing handcraft lexicons in the task of classifying extreme reviews. For the purpose of evaluation, we used five different datasets. Before defining the evaluation protocol and showing the results, we describe the resources, both lexicons and corpus-based datasets, used in the experiments.

Algorithm 1 Identifying very negative

```
1: docScore = 0
2: for Each document do
3:    WordScore = 0
4:    for Each Word in document do
5:       if Word IN VN then
6:          WordScore= -1 {the word in VN ( very negative words).}
7:       else
8:          WordScore= +1 {the word in NVN ( not very negative words).}
9:       end ifdocScore = docScore + WordScore
10:   end for
11:   if docScore <= 0 then
12:      docScore= Very Negative
13:   else
14:      docScore= Not Very Negative
15:   end if
16:   return docScore
17: end for
```

Figure 6.1: Algorithm to assign very negative classification to an input document.

Algorithm 2 Identifying very positive

```
docScore = 0
for Each document do
    WordScore = 0
    for Each Word in document do
        if Word IN VP then
            WordScore= +1 {the word in VP ( very positive words).}
        else
            WordScore= -1 {the word in NVP ( not very positive words).}
        end ifdocScore = docScore + WordScore
    end for
    if docScore >= 0 then
        docScore= Very Positive
    else
        docScore= Not Very Positive
    end if
    return docScore
end for
```

Figure 6.2: Algorithm to assign very positive classification to an input document.

6.2 The evaluated lexicons

As mentioned earlier, there are many popular and available sentiment lexicons. However, for the purpose of comparison, we need lexicons with properties according to the following two criteria:

– First, every entry in the dictionary is required to be assigned a PoS tag.

– Second, every entry must be associated with a score according to its polarity strength.

Four lexicons will be compared: the two lexicons we built using our strategy, called VERY-NEG, VERY-POS, the handcrafted lexical resource reported in Taboada et al. (2011), called SO-CAL, and SentiWords (L. et al., 2015).

In order to ensure that all cases are tested, we created lexicons at two different borderline (B) values: B=1 and B=2. we have already described in Chapter 3.

Each of our two lexicons, VERY-NEG and VERY-POS, consists of two lists derived from different values of B, as shown in Tables 6.1 and 6.2.

In order to compare the lexicons, SO-CAL and SentiWords were prepared in the same way as VERY-NEG and VERY-POS.

As far as SentiWords was concerned, we modified the range of values in order to make it similar to that of SO-CAL, making the two lexicons comparable. For this purpose, we multiplied polarity scores by 5 to provide polarity values within the -5 to 5 range, instead of -1 to 1, exactly in the same way as has been done in Taboada et al. (2011).

To make sure that the comparison of the performance of the lexicons will be fair, SO-CAL and SentiWords were divided into several lexicons. More precisely, they were split into two scales, Negative Polarity (NP) and Positive Polarity (PP), with four partitions on each scale, according to the polarity scores. The different lexicons derived from the original SO-CAL and SentiWords are defined as follows:

- **NP1:** The VN class consists of the words that are ranked as -4 and -5. The other class (NVN) contains the rest of the words.
- **NP2:** VN consists of the words that are rated as -3, -4 and -5. NVN contains the rest of the words.
- **NP3:** VN consists of the words that carry all negative ranks except -1, while the rest were considered as belonging to the class NVN.
- **NP4:** VN class consists of words with all negative ranks from -5 to -1, while NVN class contains all the words from positive ranks: from +1 to +5.
- **PP1:** The VP class consists of the words that are ranked as +4 and +5. The second class (NVP) contains the rest of the words.
- **PP2:** VP consists of the words that are rated as +3, +4 and +5. NVP contains the rest of the words.
- **PP3:** VP consists of the words that carry all positive ranks except +1, while the rest were considered as belonging to the NVP class.
- **PP4:** VP class consists of words with all positive ranks (from +5 to +1), while NVP class contains all the words with negative ranks: from -1 to -5.

Tables 6.1 and 6.2 show the total number of words of all the evaluated partitions of lexicons. The tables also include the number of words of each lexicon partition for each class (VN, NVN, VP, NVP).

	Number of words			VN			NVN		
Lexicon	ADJ	ADV	Total	ADJ	ADV	Total	ADJ	ADV	Total
VERY-NEG B=1	11670	2790	14460	4178	1092	5270	7492	1698	9190
VERY-NEG B=2	11557	2771	14328	4966	1266	6232	6591	1505	8096
SO-CAL NP1	2826	876	3702	189	62	251	2637	814	3451
SO-CAL NP2	2826	876	3702	536	135	671	2290	741	3031
SO-CAL NP3	2826	876	3702	1080	289	1369	1746	587	2333
SO-CAL NP4	2826	876	3702	1576	429	2005	1250	447	1697
SentiWords NP1	13425	2811	16236	156	4	160	13269	2807	16076
SentiWords NP2	13425	2811	16236	1132	24	1156	12293	2787	15080
SentiWords NP3	13425	2811	16236	4016	189	4205	9409	2622	12031
SentiWords NP4	13425	2811	16236	7612	540	8152	5813	2271	8084

Table 6.1: Negative lexicons: total number of words (adjectives and adverbs) for each lexicon, and number of words for each class (VN and NVN) in each lexicon

	Number of words			VP			NVP		
Lexicon	ADJ	ADV	total	ADJ	ADV	Total	ADJ	ADV	Total
VERY-POS B=1	11402	2769	14171	4721	1163	5884	6681	1606	8287
VERY-POS B=2	11472	2772	14244	5753	1339	7092	5719	1433	7152
SO-CAL PP1	2826	876	3702	239	75	314	2587	801	3388
SO-CAL PP2	2826	876	3702	512	167	679	2314	709	3023
SO-CAL PP3	2826	876	3702	835	292	1127	2155	628	2783
SO-CAL PP4	2826	876	3702	1250	447	1697	1576	429	2005
SentiWords NP1	13425	2811	16236	130	13	143	13295	2798	16093
SentiWords NP2	13425	2811	16236	581	34	615	12844	2777	15621
SentiWords NP3	13425	2811	16236	2418	250	2668	11007	2561	13568
SentiWords NP4	13425	2811	16236	5813	2271	8084	7612	540	8152

Table 6.2: Positive lexicons: total number of words (adjectives and adverbs) for each lexicon, and number of words for each class (VP and NVP) in each lexicon

Evaluation

The lexicons are evaluated on the five collections of scaled reviews by using the classification algorithm explained above in Figs 6.1 and 6.3.

Equation 6.1 defines precision P_{neg}, which is applied to evaluate the classification VN *Vs.* NVN. Similarly, Equation 6.2 defines precision P_{pos}, which is applied to VP *vs.* NVP classification.

$$P_{neg} = \frac{trueVN}{trueVN + falseVN} \tag{6.1}$$

$$P_{neg} = \frac{trueVP}{trueVP + falseVP} \tag{6.2}$$

Equation 6.3 defines recall R_{neg}, used for VN *Vs.* NVN classification. Equation 6.4 defines recall R_{pos}, for VP *Vs.* NVP

$$R_{neg} = \frac{trueVN}{trueVN + falseNVN} \tag{6.3}$$

$$R_{pos} = \frac{trueVP}{trueVP + falseNVP} \tag{6.4}$$

Equations 6.5 and 6.6 are used to compute the F-score, which is the weighted average of the precision and recall.

$$F1_{neg} = 2 * \frac{P_{neg} * R_{neg}}{P_{neg} + R_{neg}} \tag{6.5}$$

$$F1_{pos} = 2 * \frac{P_{pos} * R_{pos}}{P_{pos} + R_{pos}} \tag{6.6}$$

6.3 Very Negative Classification (VN vs NVN)

Tables 6.3, 6.4 and 6.5 show the scores (in terms of (P_{neg}, R_{neg}, and $F1_{neg}$) of the VN and NVN classes for the three lexicons across the four partitions. The experiments were carried out by applying the algorithm described in Fig 6.1. Tables 6.3 and 6.4 summarize the results using the SO-CAL and SentiWords lexicons in all partitions (NP1,NP2, NP3 and NP4). The most interesting finding is that the best $F1_{neg}$ has been achieved when using partition NP4 in both lexicons. Table 6.5 summarizes the results using two versions of our lexicon: the first lexicon was built with borderline value $B = 1$, and the second one with $B = 2$.

	NP1			NP2			NP3			NP4		
Dataset	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$
Books	0.36	0.06	0.10	0.47	0.13	0.20	0.50	0.26	0.34	0.46	0.50	0.48
DVDs	0.60	0.10	0.17	0.58	0.18	0.28	0.56	0.31	0.40	0.48	0.51	0.49
Electronics	0.57	0.13	0.21	0.62	0.20	0.31	0.62	0.29	0.39	0.55	0.49	0.52
Kitchens	0.59	0.10	0.17	0.64	0.19	0.29	0.66	0.29	0.40	0.57	0.48	0.52
Movies	0.13	0.03	0.05	0.30	0.14	0.19	0.40	0.30	0.34	0.42	0.55	0.48

Table 6.3: Polarity classification results for all collections with the SO-CAL lexicon, in terms of Precision (P_{neg}), Recall (R_{neg}) and $F1_{neg}$ scores for very negative (VN) and other (NVN) class of documents. The best $F1_{neg}$ for the VN class in each dataset is highlighted (in bold).

	NP1			NP2			NP3			NP4		
Dataset	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$
Books	0.42	0.01	0.02	0.35	0.01	0.03	0.28	0.04	0.07	0.24	0.43	0.31
DVDs	0.33	0.01	0.01	0.53	0.03	0.06	0.58	0.13	0.22	0.49	0.41	0.44
Electronics	0.26	0.01	0.01	0.37	0.02	0.03	0.63	0.18	0.28	0.57	0.49	0.53
Kitchens	0.36	0.01	0.01	0.56	0.01	0.03	0.71	0.17	0.27	0.62	0.45	0.52
Movies	0.09	0.00	0.00	0.31	0.01	0.01	0.32	0.05	0.08	0.44	0.25	0.32

Table 6.4: Polarity classification results for all collections with the SentiWords lexicon, in terms of Precision (P_{neg}), Recall (R_{neg}) and $F1_{neg}$ scores for very negative (VN) and other (NVN) documents. The best $F1_{neg}$ for the VN class in each dataset is highlighted (in bold).

	VERY-NEG B=1			VERY-NEG B=2		
Dataset	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$
Books	0.42	0.64	0.51	0.40	0.80	**0.53**
DVDs	0.43	0.76	**0.55**	0.88	0.88	0.53
Electronics	0.50	0.80	**0.62**	0.45	0.86	0.59
Kitchen	0.52	0.70	**0.60**	0.47	0.80	0.59
Movies	0.42	0.77	**0.54**	0.39	0.89	**0.54**

Table 6.5: Polarity classification results for all collections with VERY-NEG lexicon, in terms of Precision (P_{neg}), Recall (R_{neg}) and $F1_{neg}$ scores for very negative (VN) and other (NVN) documents. The best $F1_{neg}$ for the VN class in each dataset is highlighted (in bold).

By comparing the results shown in the three tables (6.3,6.4 and 6.5) on the three lexicons, we may make the following observations:

– The best $F1_{neg}$ scores in all datasets have been achieved by the two versions of VERY-NEG lexicon. The $B = 1$ version is the best on DVDs, Electronics and Kitchen datasets, while the $B = 2$ version performs better on Books and Movies.

– In all tests, we can observe that the evaluation values for identifying the VN class are low.

– We can also observe in all tests that the best $F1_{neg}$ scores were reached using the Electronics and Kitchen datasets, while the worst values were obtained with Movies and Books.

– In general, the behavior of Movies and Books tends to be different from the other datasets.

– The lexicon we proposed, VERY-NEG, consistently outperforms the other lexicons on the five datasets as shown in Fig 6.3.

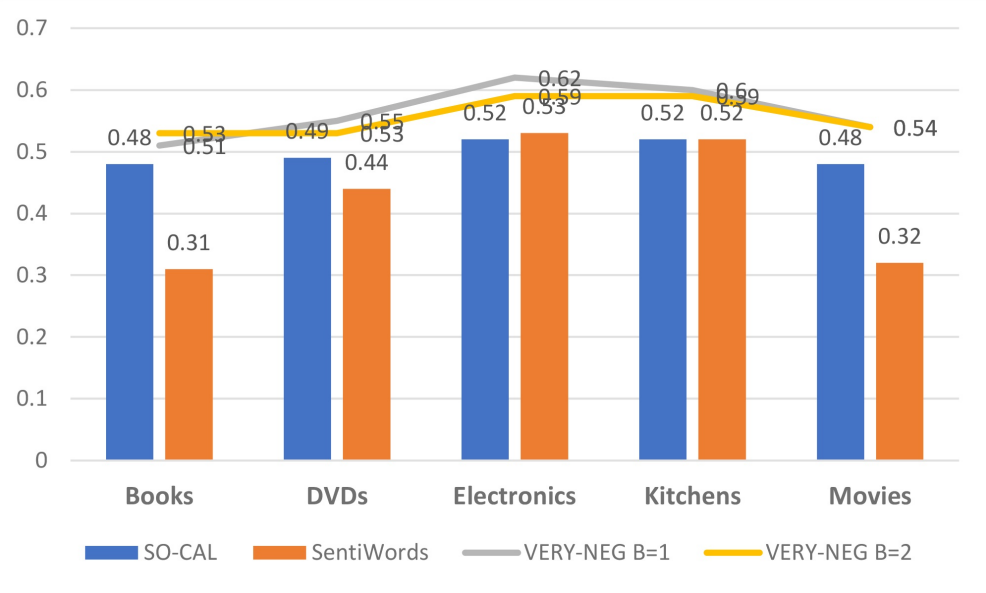

Figure 6.3: The best performance ($F1_{neg}$) obtained by all lexicons on all datasets for identifying very negative documents (VN *vs* NVN).

6.4 Very Positive Classification (VP vs NVP)

Tables 6.6, 6.7, and 6.8 show the scores (in terms of (P_{pos}, R_{pos}, and $F1_{pos}$) of VP/NVP for the three lexicons across the four partitions. The experiments were carried out by applying the algorithm described above in Fig 6.3. Tables 6.6 and 6.7 show the results obtained using the SO-CAL and SentiWords lexicons. The best $F1_{pos}$ scores in both lexicons on all datasets were achieved when partition PP4 was used. Table 6.8 summarizes the results using two versions of our lexicon again: the one defined with $B = 1$, and the second one with $B = 2$.

	PP1			PP2			PP3			PP4		
Dataset	P_{pos}	R_{pos}	$F1_{pos}$	P_{pos}	R_{pos}	$F1_{pos}$	P_{pos}	R_{pos}	$F1_{pos}$	P_{pos}	R_{pos}	$F1_{pos}$
Books	0.61	0.17	0.27	0.54	0.34	0.42	0.52	0.55	0.53	0.41	0.94	0.57
DVDs	0.66	0.21	0.32	0.58	0.38	0.46	0.54	0.56	0.55	0.41	0.95	0.58
Electronics	0.54	0.26	0.35	0.51	0.40	0.45	0.49	0.60	0.54	0.38	0.94	0.54
Kitchens	0.53	0.23	0.32	0.53	0.36	0.43	0.50	0.55	0.52	0.42	0.97	0.59
Movies	0.75	0.11	0.20	0.60	0.29	0.39	0.52	0.49	0.50	0.35	0.94	0.51

Table 6.6: Polarity classification results for all collections with SO-CAL lexicon, in terms of Precision (P_{pos}), Recall (R_{pos}) and $F1_{pos}$ scores for very positive (VP) and other (NVP) documents. The best $F1_{pos}$ for the VP class in each dataset is highlighted (in bold).

	PP1			PP2			PP3					
Dataset	P_{pos}	R_{pos}	$F1_{pos}$	P_{pos}	R_{pos}	$F1_{pos}$	P_{pos}	R_{pos}	$F1_{pos}$	P_{pos}	R_{pos}	$F1_{pos}$
Books	0.76	0.06	0.12	0.66	0.13	0.22	0.60	0.38	0.46	0.40	0.93	0.55
DVDs	0.65	0.07	0.21	0.64	0.13	0.22	0.59	0.38	0.46	0.39	0.92	0.55
Electronics	0.70	0.11	0.19	0.71	0.19	0.30	0.63	0.41	0.50	0.40	0.93	0.55
Kitchens	0.61	0.07	0.13	0.63	0.17	0.27	0.65	0.37	0.47	0.43	0.94	0.59
Movies	0.64	0.01	0.03	0.63	0.05	0.09	0.55	0.27	0.36	0.31	0.95	0.47

Table 6.7: Polarity classification results for all collections with SO-CAL lexicon, in terms of Precision (P_{pos}), Recall (R_{pos}) and $F1_{pos}$ scores for most positive (VP) and other (NVP) documents. The best $F1_{pos}$ for the VP class in each dataset is highlighted (in bold).

By comparing the results to differentiate between VP and NVP, we may make the following observations:

– In all datasets, the highest $F1_{pos}$ values were reached by the version of VERY-POS lexicon with $B = 2$.

	VERY-POS B=1			VERY-POS B=2		
Dataset	P_{pos}	R_{pos}	$F1_{pos}$	P_{pos}	R_{pos}	$F1_{pos}$
Books	0.67	0.55	0.61	0.61	0.67	**0.64**
DVDs	0.68	0.49	0.57	0.63	0.61	**0.62**
Electronics	0.63	0.42	0.50	0.57	0.52	**0.54**
Kitchen	0.63	0.43	0.51	0.60	0.60	**0.60**
Movies	0.63	0.41	0.50	0.55	0.58	**0.57**

Table 6.8: Polarity classification results for all collections with VERY-POS lexicon, in terms of Precision (P_{pos}), Recall (R_{pos}) and $F1_{pos}$ scores for very positive (VP) and other (NVP) documents. The best $F1_{pos}$ for the VP class in each dataset is highlighted (in bold).

– The evaluation values for identifying the VP class are again low.

– Surprisingly, the highest $F1_{pos}$ values were obtained on the Books dataset while the worst scores were on Movies and Electronics. This was not expected because the Electronics dataset was the dataset with the highest scores in identifying the most negative views and the Books was the dataset with the lowest scores.

– The lexicon we proposed, VERY-POS, consistently outperforms the other lexicons on the five datasets as shown in Fig 6.4.

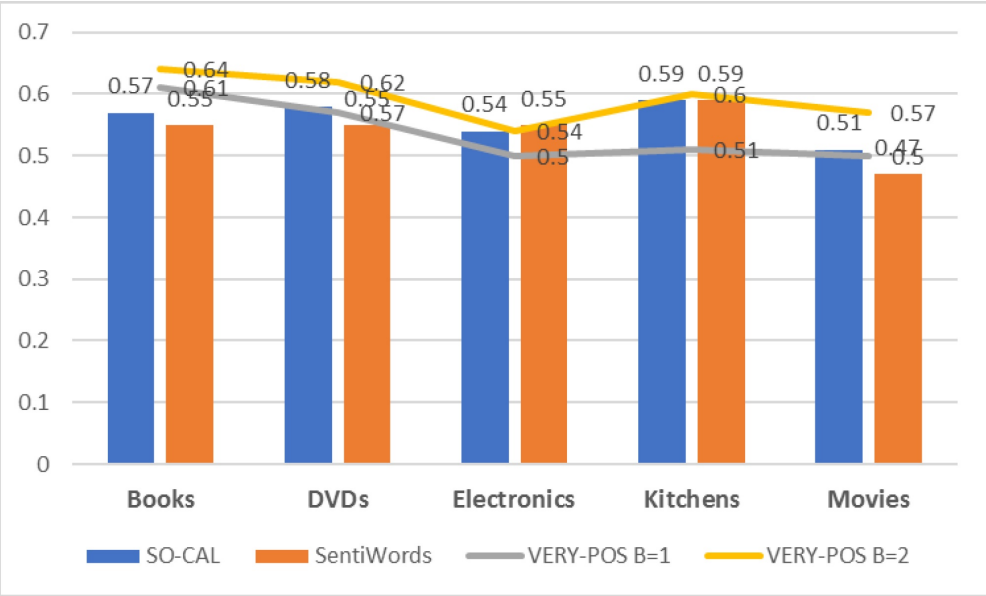

Figure 6.4: The best performance ($F1_{pos}$) obtained by all lexicons on all datasets for identifying the most positive documents.

Discussion

The experiments carried out show that our automatic strategy for building corpus-based lexicons improves existing manual resources for the task of identifying extreme opinions.

The low values achieved by the sentiment classification method can be partially explained by the difficulty of the task. The difference between extreme and not extreme is a subjective continuum without clearly defined edges. It is much more difficult to grasp this difference than that between negative and positive views. Notice that there is a well established barrier between positive and negative values consisting of neutral words. By contrast, no qualitative borderline can be found between very negative and less negative scores or very positive and less positive scores.

The poor results with the Movies dataset might be due to the fact that films are symbolic objects with an internal plot and, thus, it is natural that a person has a very positive opinion of a plot with many negative elements. The same is true the other way round. This makes the sentiment analysis of movies a very difficult task. As books are also symbolic objects, we are

not able to explain why the results of Books do not follow the same tendency as Movies in the VN/NVN task.

On the other hand, a possible explanation for the very poor performance of SO-CAL and SentiWords lexicons in the first three partitions (NP1, NP2, NP3, PP1, PP2, and PP3) might be the unbalanced number of words across the two classes in each case as shown in tables 6.1 and 6.2.

Thanks to the detailed observation of the results, one interesting finding is the relevance of the borderline calibration feature (B=1, B=2) provided by our proposed method to create sentiment lexicons. This feature gives an interesting flexibility to deal with sentiment analysis tasks. This is clearly exhibited by the difference in performance in each classification task, as there are tasks where the performance is better with the dictionaries at the borderline B=2, while others are better at B=1. For example, the performance of VERY-NEG/B=1 is better than VERY-NEG/B=2 in the task VN vs NVN. The opposite occurs in task VP vs NVP, where VERY-POS/B=2 performs better than VERY-POS/B=1.

Another important finding is that our lexicons have demonstrated stable performance between using supervised and unsupervised machine learning approaches while the performance of other lexicons was not stable between the two approaches, especially in the task of identifying very negative views.

Finally, the results of this study indicate that specific domain lexicons tend to give better results in sentiment classification than general dictionaries. Through the results of our experiments, we observed that the domain-specific lexicon, SPLM, which we built from a corpus of movie reviews, outperformed the other dictionaries in standard classification task. The same happened for extreme opinion classification when we built VERY-NEG and VERY-POS lexicons form corpora of restaurant and hotel reviews in subsection 5.4.2.

CHAPTER 7

CONCLUSIONS

Lexicon-based approaches are very popular in sentiment analysis and opinion mining, and they play a key role in all applications in this field. We described in this thesis a method for automatically building domain-specific polarity lexicons from annotated corpora.

A standard polarity lexicon has been built using movie reviews, and we evaluated its quality in an indirect way. More precisely, the lexicon was used to train a sentiment classifier which was evaluated by means of well-known datasets. The experiments reported in our work show that the lexicon we generated automatically outperforms other manual lexicons when they are used as features of a supervised sentiment classifier. However, our corpus-based strategy is not restricted to a particular domain. It is generic enough to be expanded to whatever domain and language if we are provided with corpora annotated in the appropriate way.

The main goal of the current thesis was to place value on extreme opinions because of their importance in various fields. For this purpose, we have adapted our learning method to automatically build a lexicon of extremely negative and positive words from labeled corpora. Then, we integrated it into a classifier to search for the extreme reviews. Our classifier identifies extreme opinions in two steps. On the one hand, it identifies extremely negative documents from the rest, and on the other, it classifies extremely positive documents from the rest.

We have measured the quality of a corpus-based sentiment lexicon and some handcrafted resources by evaluating their performance in a supervised strategy to classify extreme opinions. The results of this indirect evaluation show that the automatically built lexicon has a

stable behavior in different datasets and even improves other manually constructed resources. Also, we studied different linguistic features for a particular task in sentiment analysis. More precisely, we examined the performance of these features within supervised learning methods (using Support Vector Machine (SVM)), to identify extreme opinions on a dataset of hotel reviews. The experiments we carried out showed that n-gram models are difficult to outperform, but we found two features that consistently outperforms the baselines: neural-based embeddings and textual features. Polarity lexicons help improve the results, but their influence is moderate. It seems that polarity lexicons would be better integrated into unsupervised techniques than into supervised strategies. And this is the last experiment we carried out.

In the last experiments, we used an unsupervised approach to search for extreme opinions. Our unsupervised classifier identifies extreme opinions in two steps as in the previous experiments. On the one hand, it identifies extremely negative documents from the rest, and on the other, it classifies extremely positive documents from the rest. Our classification algorithm is based on a very basic word-matching scheme to carried out unsupervised sentiment analysis.

Our automatically built lexicons have been compared with handcrafted lexicons, by taking into account some partitions of them. For this purpose, we divide each handcrafted lexicon into partitions depending on the polarity weight of each word. Then, the experiments were carried out on each partition separately.

The results of the experiments show that our lexicons are better suited to identify the extreme opinions than two well-known resources, namely SO-CALL and SentiWords (a version of SentiWordNet).

Finally, the main conclusion that we can draw from all the experiments carried out is the following: our automatically built lexicons have a stable behavior in their use with both supervised and unsupervised machine learning approaches, while the performance of other existing lexicons was not so stable and reliable when applied to the two approaches, especially in the task of identifying very negative views.

Bibliography

Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12.

Abbasi, A., France, S., Zhang, Z., and Chen, H. (2011). Selecting attributes for sentiment classification using feature relation networks. *IEEE Transactions on Knowledge and Data Engineering*, 23(3):447–462.

Abdul-Mageed, M., Diab, M., and Kübler, S. (2014). Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.

Aly, M. and Atiya, A. (2013). Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 494–498.

Appel, O., Chiclana, F., Carter, J., and Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems*, 108:110–124.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Balahur, A., Hermida, J. M., and Montoyo, A. (2012). Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE Trans. Affective Computing*, 3(1):88–101.

Basari, A. S. H., Hussin, B., Ananta, I. G. P., and Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, 53:453–462.

Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., and Subrahmanian, V. S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *ICWSM*. Citeseer.

Benamara, F., Chardon, B., Mathieu, Y., and Popescu, V. (2011). Towards context-based subjectivity analysis. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1180–1188.

Bilal, M., Israr, H., Shahid, M., and Khan, A. (2016). Sentiment classification of roman-urdu opinions using naïve bayesian, decision tree and knn classification techniques. *Journal of King Saud University-Computer and Information Sciences*, 28(3):330–344.

Blitzer, J., Dredze, M., and Pereira, F. (2007a). Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Blitzer, J., Dredze, M., Pereira, F., et al. (2007b). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447.

Boiy, E. and Moens, M.-F. (2009). A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval*, 12(5):526–558.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Bosco, C., Patti, V., and Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer.

Cambria, E. (2013). An introduction to concept-level sentiment analysis. In *Mexican international conference on artificial intelligence*, pages 478–483. Springer.

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Cambria, E., Gastaldo, P., Bisio, F., and Zunino, R. (2015). An elm-based model for affective analogical reasoning. *Neurocomputing*, 149:443–455.

Cambria, E., Havasi, C., and Hussain, A. (2012). Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *FLAIRS conference*, pages 202–207.

Cambria, E., Olsher, D., and Rajagopal, D. (2014). Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence*.

Cambria, E., Poria, S., Bajpai, R., and Schuller, B. W. (2016). Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *COLING*, pages 2666–2677.

Cambria, E., Poria, S., Hazarika, D., and Kwok, K. (2018). Senticnet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of AAAI*.

Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*, volume 10.

Carrillo de Albornoz, J., Plaza, L., and Gervás, P. (2010). A hybrid approach to emotional sentence polarity and intensity classification. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 153–161. Association for Computational Linguistics.

Chao, A. F. and Yang, H.-L. (2018). Using chinese radical parts for sentiment analysis and domain-dependent seed set extraction. *Computer Speech & Language*, 47:194–213.

Chenlo, J. M. and Losada, D. E. (2014). An empirical study of sentence features for subjectivity and polarity classification. *Information Sciences*, 280:275–288.

Chevalier, J. A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.

Chinsha, T. and Joseph, S. (2015). A syntactic approach for aspect based opinion mining. In *2015 IEEE International Conference on Semantic Computing (ICSC)*, pages 24–31. IEEE.

Cho, H., Kim, S., Lee, J., and Lee, J.-S. (2014). Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. *Knowledge-Based Systems*, 71:61–71.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

comScore, I. and Kelsey, T. (2007). Online consumer-generated reviews have significant impact on offline purchase behavior. *Retrieved*, 5(16):2011.

Coussement, K. and Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, 36(3):6127–6134.

Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J., and Vallejo, C. G. (2013). 'long autonomy or long delay?'the importance of domain in opinion mining. *Expert Systems with Applications*, 40(8):3174–3184.

Cruz, F. L., Troyano, J. A., Pontes, B., and Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Dai, A. M., Olah, C., and Le, Q. V. (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

Dang, Y., Zhang, Y., and Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4):46–53.

Desmet, B. and Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358.

Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM.

Duwairi, R. M. and Qarqaz, I. (2014). Arabic sentiment analysis using supervised classification. In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, pages 579–583. IEEE.

Esuli, A. and Sebastiani, F. (2007). Sentiwordnet: A high-coverage lexical resource for opinion mining. *Evaluation*, pages 1–26.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

Fellbaum, C. (1998). A semantic network of English: The mother of all WordNets. *Computer and the Humanities*, 32:209–220.

Garcia-Moya, L., Anaya-Sanchez, H., and Berlanga-Llavori, R. (2013). Retrieving product features and opinions from customer reviews. *IEEE Intelligent Systems*, 28(3):19–27.

Gatti, L., Guerini, M., and Turchi, M. (2016). Sentiwords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421.

Gerani, S., Carman, M. J., and Crestani, F. (2009). Investigating learning approaches for blog post opinion retrieval. In *European Conference on Information Retrieval*, pages 313–324. Springer.

Ghiassi, M. and Saidane, H. (2005). A dynamic architecture for artificial neural networks. *Neurocomputing*, 63:397–413.

Ghiassi, M., Skinner, J., and Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282.

Habernal, I., Ptáček, T., and Steinberger, J. (2015). Reprint of "supervised sentiment analysis in czech social media". *Information Processing & Management*, 51(4):532–546.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181. Association for Computational Linguistics.

Hemmatian, F. and Sohrabi, M. K. (2017). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, pages 1–51.

Hiroshi, K., Tetsuya, N., and Hideo, W. (2004). Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th international conference on Computational Linguistics*, page 494. Association for Computational Linguistics.

Horrigan, J. (2008). *Online shopping: Internet users like the convenience but worry about the security of their financial information*. Pew Internet & American Life Project Washington, DC.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Huang, S., Niu, Z., and Shi, C. (2014). Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems*, 56:191–200.

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Jeyapriya, A. and Selvi, C. K. (2015). Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*, pages 548–552. IEEE.

Jiménez-Zafra, S. M., Martín-Valdivia, M. T., Martínez-Cámara, E., and Ureña-López, L. A. (2016). Combining resources to improve unsupervised sentiment analysis at aspect-level. *Journal of Information Science*, 42(2):213–229.

Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 219–230. ACM.

Jo, Y. and Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Joshi, A., Bhattacharyya, P., and Ahire, S. (2017). Sentiment resources: Lexicons and datasets. In *A Practical Guide to Sentiment Analysis*, pages 85–106. Springer.

Kamps, J., Marx, M., Mokken, R. J., De Rijke, M., et al. (2004). Using wordnet to measure semantic orientations of adjectives. In *LREC*, volume 4, pages 1115–1118. Citeseer.

Kanayama, H. and Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 355–363. Association for Computational Linguistics.

Kaur, A. and Gupta, V. (2013). A survey on sentiment analysis and opinion mining techniques. *Journal of Emerging Technologies in Web Intelligence*, 5(4):367–371.

Kennedy, A. and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125.

Khan, F. H., Bashir, S., and Qamar, U. (2014). Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57:245–257.

Kim, S.-M. and Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8. Association for Computational Linguistics.

Kouloumpis, E., Wilson, T., and Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164.

Kranjc, J., Smailović, J., Podpečan, V., Grčar, M., Žnidaršič, M., and Lavrač, N. (2015). Active learning for sentiment analysis on data streams: Methodology and workflow

implementation in the clowdflows platform. *Information Processing & Management*, 51(2):187–203.

L., G., Guerini, M., and Turchi, M. (2015). Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 99.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Lee, J., Park, D.-H., and Han, I. (2008). The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic commerce research and applications*, 7(3):341–352.

Li, S.-K., Guan, Z., Tang, L.-Y., and Chen, Z. (2012). Exploiting consumer reviews for product feature ranking. *Journal of Computer Science and Technology*, 27(3):635–649.

Li, S.-T. and Tsai, F.-C. (2013). A fuzzy conceptualization model for text mining with application in opinion polarity classification. *Knowledge-Based Systems*, 39:23–33.

Liao, C., Feng, C., Yang, S., and Huang, H.-Y. (2016). A hybrid method of domain lexicon construction for opinion targets extraction using syntax and semantics. *Journal of Computer Science and Technology*, 31(3):595–603.

Lin, Z., Tan, S., Liu, Y., Cheng, X., and Xu, X. (2013). Cross-language opinion lexicon extraction using mutual-reinforcement label propagation. *PloS one*, 8(11):e79294.

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Liu, L., Nie, X., and Wang, H. (2012). Toward a fuzzy domain sentiment ontology tree for sentiment analysis. In *Image and Signal Processing (CISP), 2012 5th International Congress on*, pages 1620–1624. IEEE.

Lpez Condori, R. E. and Salgueiro Pardo, T. A. (2017). Opinion summarization methods. *Expert Systems with Applications: An International Journal*, 78(C):124–134.

Lu, Y., Castellanos, M., Dayal, U., and Zhai, C. (2011). Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM.

Lu, Y., Zhai, C., and Sundaresan, N. (2009). Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*, pages 131–140. ACM.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Maks, I. and Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53(4):680–688.

Martín-Valdivia, M.-T., Martínez-Cámara, E., Perea-Ortega, J.-M., and Ureña-López, L. A. (2013). Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10):3934 – 3942.

McDonald, R., Hannan, K., Neylon, T., Wells, M., and Reynar, J. (2007). Structured models for fine-to-coarse sentiment analysis. In *Annual meeting-association for computational linguistics*, volume 45, page 432.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Meng, X. and Wang, H. (2009). Mining user reviews: from specification to summarization. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 177–180. Association for Computational Linguistics.

Miao, Q., Li, Q., and Dai, R. (2009). Amazing: A sentiment mining and retrieval system. *Expert Systems with Applications*, 36(3):7192–7198.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mohammad, S. M. and Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Mohammad, S. M. and Turney, P. D. (2013a). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Mohammad, S. M. and Turney, P. D. (2013b). Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., and Ureña-López, L. A. (2015). A spanish semantic orientation approach to domain adaptation for polarity classification. *Information Processing & Management*, 51(4):520–531.

Moraes, R., Valiati, J. F., and Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications*, 40(2):621–633.

Moreo, A., Romero, M., Castro, J., and Zurita, J. M. (2012). Lexicon-based comments-oriented news sentiment analyzer system. *Expert Systems with Applications*, 39(10):9166–9180.

Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., and Ghosh, R. (2013). Spotting opinion spammers using behavioral footprints. In *Proceedings of the*

19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 632–640. ACM.

Mukherjee, A., Liu, B., and Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200. ACM.

Mukherjee, A., Liu, B., Wang, J., Glance, N., and Jindal, N. (2011). Detecting group review spam. In *Proceedings of the 20th international conference companion on World wide web*, pages 93–94. ACM.

Mullen, T. and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Narayanan, R., Liu, B., and Choudhary, A. (2009). Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 180–189. Association for Computational Linguistics.

Nasukawa, T. and Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77. ACM.

Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Nielsen, F. Å., Hansen, L. K., Arvidsson, A., and Colleoni, E. (2011). Finn årup nielsen abstract afinn is a list of english words rated for valence with an integer between minus five (negative) and plus five (positive). the words have been manually labeled by finn årup nielsen in 2009-2011. the file is tab-separated. there are two versions.

Nishikawa, H., Hasegawa, T., Matsuo, Y., and Kikui, G. (2010). Opinion summarization with integer linear programming formulation for sentence extraction and ordering. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 910–918. Association for Computational Linguistics.

Ott, M., Cardie, C., and Hancock, J. (2012). Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web*, pages 201–210. ACM.

Paltoglou, G. and Thelwall, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1386–1395. Association for Computational Linguistics.

Pang, B. and Lee, L. (2004a). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Pang, B. and Lee, L. (2004b). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Penalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodriguez-Garcia, M. A., Moreno, V., Fraga, A., and Sanchez-Cervantes, J. L. (2014). Feature-based opinion mining through ontologies. *Expert Systems with Applications*, 41(13):5995–6008.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Pham, D.-H. and Le, A.-C. (2018). Learning multiple layers of knowledge representation for aspect based sentiment analysis. *Data & Knowledge Engineering*, 114:26–39.

Poria, S., Cambria, E., Winterstein, G., and Huang, G.-B. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 69:45–63.

Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D., and Bandyopadhyay, S. (2013). Enhanced senticnet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38.

Potts, C. (2010). On the negativity of negation. In *Semantics and Linguistic Theory*, volume 20, pages 636–659.

Potts, C. (2011). Developing adjective scales from user-supplied textual metadata. In *NSF Workshop on Restructuring Adjectives in WordNet. Arlington, VA*.

Prabowo, R. and Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157.

Quan, C. and Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences*, 272:16–28.

Rathan, M., Hulipalled, V. R., Venugopal, K., and Patnaik, L. (2018). Consumer insight mining: aspect based twitter opinion mining of mobile phone reviews. *Applied Soft Computing*, 68:765–773.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112. Association for Computational Linguistics.

Rustamov, S., Mustafayev, E., and Clements, M. (2013). Sentence-level subjectivity detection using neuro-fuzzy models. In *Proceedings of the 4th Workshop on*

Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 108–114.

Saleh, M. R., Martín-Valdivia, M. T., Montejo-Ráez, A., and Ureña-López, L. (2011). Experiments with svm to classify opinions in different domains. *Expert Systems with Applications*, 38(12):14799–14804.

Seki, Y., Kando, N., and Aono, M. (2009). Multilingual opinion holder identification using author and authority viewpoints. *Information Processing & Management*, 45(2):189–199.

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., and Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311:18–38.

Severyn, A., Moschitti, A., Uryupina, O., Plank, B., and Filippova, K. (2016). Multi-lingual opinion mining on youtube. *Information Processing & Management*, 52(1):46–60.

Shah, R. R., Yu, Y., Verma, A., Tang, S., Shaikh, A. D., and Zimmermann, R. (2016). Leveraging multimodal information for event summarization and concept-level sentiment analysis. *Knowledge-Based Systems*, 108:102–109.

Sharma, R., Nigam, S., and Jain, R. (2014). Opinion mining of movie reviews at document level. *arXiv preprint arXiv:1408.3829*.

Silva, N. F. F. D., Coletta, L. F., and Hruschka, E. R. (2016). A survey and comparative study of tweet sentiment analysis via semi-supervised learning. *ACM Computing Surveys (CSUR)*, 49(1):15.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Taboada, M., Anthony, C., and Voll, K. (2006). Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 427–432.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Takamura, H., Inui, T., and Okumura, M. (2007). Extracting semantic orientations of phrases from dictionary. In *HLT-NAACL*, volume 2007, pages 292–299.

Tang, H., Tan, S., and Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7):10760–10773.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology*, 61(12):2544–2558.

Tripathy, A., Agrawal, A., and Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 57:117–126.

Tsai, A., Tsai, R. T.-H., and Hsu, J. Y.-j. (2013). Building a concept-level sentiment dictionary based on commonsense knowledge. *IEEE Intelligent Systems*, page 1.

Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Wang, D., Zhu, S., and Li, T. (2013). Sumview: A web-based engine for summarizing product reviews and customer opinions. *Expert Systems with Applications*, 40(1):27–33.

Wang, G., Xie, S., Liu, B., and Yu, P. S. (2012). Identify online store review spammers via social review graph. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):61.

Wang, S., Li, D., Song, X., Wei, Y., and Li, H. (2011). A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications*, 38(7):8696–8702.

Weichselbraun, A., Gindl, S., and Scharl, A. (2013). Extracting and grounding contextualized sentiment lexicons. *IEEE Intelligent Systems*, (2):39–46.

Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497. Springer.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Xia, R., Zong, C., and Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6):1138–1152.

Yu, Y., Duan, W., and Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4):919–926.

Zhai, Z., Liu, B., Wang, J., Xu, H., and Jia, P. (2012). Product feature grouping for opinion mining. *IEEE Intelligent Systems*, 27(4):37–44.

Zhang, Z. and Singh, M. P. (2014). Renew: A semi-supervised framework for generating domain-specific lexicons and sentiment analysis. In *ACL (1)*, pages 542–551.

Zhang, Z., Ye, Q., Zhang, Z., and Li, Y. (2011). Sentiment classification of internet restaurant reviews written in cantonese. *Expert Systems with Applications*, 38(6):7674–7682.

Zhu, J., Wang, H., Zhu, M., Tsou, B. K., and Ma, M. (2011). Aspect-based opinion polling from customer reviews. *IEEE Transactions on Affective Computing*, 2(1):37–49.

List of Figures

List of Tables

Appendix A

Publications

– (JCR Journal) Sattam Almatarneh, Pablo Gamallo. A lexicon based method to search for extreme opinions, PloS one (Impact factor in 2016: 2.806 (Q1)), 2018. Studies in sentiment analysis and opinion mining have been focused on many aspects related to opinions, namely polarity classification by making use of positive, negative or neutral values. However, most studies have overlooked the identification of extreme opinions (most negative and most positive opinions) in spite of their vast significance in many applications. We use an unsupervised approach to search for extreme opinions, which is based on the automatic construction of a new lexicon containing the most negative and most positive words.
https://doi.org/10.1371/journal.pone.0197816

– (LNCS) Sattam Almatarneh, Pablo Gamallo. Linguistic Features to Identify Extreme Opinions: An Empirical Study. Proceedings of IDEAL 2018, 19th International Conference on Intelligent Data Engineering and Automated Learning, Madrid, Spain, November 2018. Lecture Notes in Computer Science volume 11314, (ISBN:978-3-030-03492-4). In this study, we combined empirical features (e.g. bag of words, word embeddings, polarity lexicons, and set of textual features) so as to identify extreme opinions and provide a comprehensive analysis of the relative importance of each set of features using hotel reviews.

– (IEEE) Sattam Almatarneh, Pablo Gamallo. A Comparative Study of Polarity Lexicons to Identify Extreme Opinions. Proceedings of SNAMS 2018, Fifth International Conference on Social Networks Analysis, the Second International Workshop on Advances in Natural Language Processing (ANLP 2018) Management

and Security, Valencia, Spain, October 2018. This paper comparing a method to automatically build a sentiment lexicon, with four well-known sentiment lexicons. For this purpose, an indirect evaluation is carried out. The lexicons are integrated into supervised sentiment classifiers and their performance is evaluated in two sentiment classification tasks in order to identify i) the most negative vs. not most negative opinions, and ii) the most positive vs. not most positive. Moreover, a set of textual features is integrated into the classifiers so as to analyze how these textual features improve the lexicon performance.

– (CCIS) Sattam Almatarneh, Pablo Gamallo. Searching for the Most Negative Opinions. Proceedings of KESW 2017, 8th International Conference on Knowledge Engineering and Semantic Web, Szczecin, Poland, November 2017. Communications in Computer and Information Science (CCIS) volume 786, (ISBN:978-3-319-69548-8). In this article, we used diversified linguistic features and supervised machine learning algorithms so as to examine their effectiveness in the process of searching for the most negative opinions.
https://doi.org/10.1007/978-3-319-69548-8_2

– (AISC) Sattam Almatarneh, Pablo Gamallo. Automatic Construction of Domain-Specific Sentiment Lexicons for Polarity Classification. Proceedings of PAAMS 2017, 8th International Conference on Practical Applications of Agents and Multi-Agent Systems, Porto, Portugal, June 2017. Advances in Intelligent Systems and Computing (AISC) volume 619, (ISBN:978-3-319-61578-3). The article describes a strategy to build sentiment lexicons (positive and negative words) from corpora. Special attention will be paid to the construction of a domain-specific lexicon from a corpus of movie reviews. Polarity words of the lexicon are assigned weights standing for different degrees of positiveness and negativeness. This lexicon is integrated into a sentiment analysis system in order to evaluate its performance in the task of sentiment classification. The experiments performed shows that the lexicon we generated automatically outperforms other manual lexicons when they are used as features of a supervised sentiment classifier.
https://doi.org/10.1007/978-3-319-61578-3_17