A COMPARATIVE STUDY OF MACHINE TRANSLATION FOR MULTILINGUAL SENTENCE-LEVEL SENTIMENT ANALYSIS

MATHEUS ARAUJO

A COMPARATIVE STUDY OF MACHINE TRANSLATION FOR MULTILINGUAL SENTENCE-LEVEL SENTIMENT ANALYSIS

Dissertação apresentada ao Programa de Pós-Graduação em Computer Science do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Computer Science.

ORIENTADOR: FABRÍCIO BENEVENUTO

Belo Horizonte
Junho de 2017

MATHEUS ARAUJO

A COMPARATIVE STUDY OF MACHINE TRANSLATION FOR MULTILINGUAL SENTENCE-LEVEL SENTIMENT ANALYSIS

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

ADVISOR: FABRÍCIO BENEVENUTO

Belo Horizonte
June 2017

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Araújo, Matheus Lima Diniz.

A663c A Comparative study of machine translation for multilingual sentence-level sentiment analysis / Matheus Lima Diniz Araújo – Belo Horizonte, 2017.
xxiv, 60 f.: il.; 29 cm.

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Fabrício Benevenuto de Souza

1. Computação – Teses. 2. Recuperação da informação. 3. Redes sociais on-line. 4. Análise de sentimentos. 5. Mineração de opinião. 6. Tradução automática. I. Orientador. II. Título.

CDU 519.6*04 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

A comparative study of machine translation for multilingual sentence-level sentiment analysis

MATHEUS LIMA DINIZ ARAÚJO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Fabrício Benevenuto
PROF. FABRÍCIO BENEVENUTO DE SOUZA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ADRIANO CÉSAR MACHADO PEREIRA
Departamento de Ciência da Computação - UFMG

Flavio Vinicius Diniz de Figueiredo
PROF. FLÁVIO VINICIUS DINIZ DE FIGUEIREDO
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 26 de junho de 2017.

I dedicate this dissertation to my parents and my little sister

Acknowledgments

There have been many people who supported me during the last two years. They have guided me, supported me and placed many opportunities in front of me. I would like to thank especially Dr. Fabrício Benevenuto, who trusted on my potential. I'll be eternally grateful for showing me the beauty in being a scientist.

I also want to thank in special my parents Celina and Virgilio for dedicating their life to give me the best they could give, and I'm sure that it was the best any son can have.

To Leticia, my smart little sister, I want to thank you for understand those days I couldn't stay at home playing with you. To my loved Larissa, thank you for being at my side even when I wasn't at yours, I love you. To Gustavo, Caju, Mari e Feliphe, to be the best friends someone can have during this journey. To the brilliant scientists at LOCUS lab who shared not just knowledge but friendship along these years. To my uncles, aunts, and cousins, for always pushing me towards my goals. To the professors at ICEX and DCC for creating this unique environment of excellence in science for Brazil.

Finally, I feel very lucky to have traced this path in my life, and I'm excited for what destiny plans for me. Therefore I thank all the mystical entities and gods responsible for it.

“There’s a difference between knowing the path and walking the path”

(Morpheus)

Resumo

Análise de sentimentos se tornou uma ferramenta chave em aplicações voltadas para mídias sociais, incluindo a classificação da opinião dos usuários sobre produtos e serviços, o monitoramento político das mídias durante campanhas eleitorais e até mesmo a influência no mercado de ações. Existem diferentes ferramentas para análise de sentimentos que exploram variadas técnicas, estas baseiam-se em dicionários léxicos ou aprendizado de máquina. Apesar do significante interesse neste tema e o grande esforço investido nesta área pela comunidade científica, quase todos os métodos existentes para análise de sentimentos foram direcionados para o contexto da língua inglesa. A maioria das estratégias para análise em diferentes línguas consiste na adaptação de um léxico já existente em inglês, sem apresentar validações ou comparações com linhas de base. Neste trabalho, realizamos uma abordagem diferente para resolver o problema de análise de sentimentos em diferentes línguas. Para isto, avaliamos 16 métodos voltados à análise de sentimentos em sentenças criadas para o inglês e os comparamos com três abordagens geradas para línguas específicas. A partir de 14 conjuntos de dados em diferentes línguas, rotulados por humanos, realizamos uma extensa análise quantitativa das abordagens criadas para múltiplos idiomas. Nossos resultados sugerem que a simples tradução automática do texto de entrada da língua específica para o inglês e, em seguida, a utilização dos métodos estado-da-arte criados para o inglês pode ser melhor que os métodos existentes desenvolvidos para uma língua específica. Nós também classificamos os métodos de acordo com sua capacidade de predição e identificamos aqueles métodos que alcançam os melhores resultados utilizando tradução automática entre as diferentes línguas. Como contribuição final para a comunidade acadêmica, compartilhamos os códigos, conjuntos de dados e o sistema iFeel 3.0, um arcabouço para análise de sentimentos em sentenças para múltiplas línguas. Esperamos que nossa metodologia se torne uma linha de base para o desenvolvimento de novos métodos de análise de sentimentos ao nível de sentenças em múltiplas línguas.

Palavras-chave: Análise de Sentimentos, Multilíngue , Tradução Automática, Redes Sociais Online, Mineração de Opinião.

Abstract

Sentiment analysis has become a key tool for several social media applications, including analysis of user's opinions about products and services, support for politics during campaigns and even for market trending. Multiple existing sentiment analysis methods explore different techniques, usually relying on lexical resources or learning approaches. Despite the significant interest in this theme and amount of research efforts in the field, almost all existing methods are designed to work with only English content. Most current strategies in many languages consist of adapting existing lexical resources, without presenting proper validations and basic baseline comparisons. In this work, we take a different step into this field. We focus on evaluating existing efforts proposed to do language specific sentiment analysis with a simple yet effective baseline approach. To do it, we evaluated sixteen methods for sentence-level sentiment analysis proposed for English, comparing them with three language-specific methods. Based on fourteen human labeled language-specific datasets, we provide an extensive quantitative analysis of existing multi-language approaches. Our primary results suggest that simply translating the input text on a specific language to English and then using one of the existing best methods developed to English can be better than the existing language specific efforts evaluated. We also rank methods according to their prediction performance and we identified the methods that acquired the best results using machine translation across different languages. As a final contribution to the research community, we release our codes, datasets, and the iFeel 3.0 system, a web framework for multilingual sentence-level sentiment analysis. We hope our system setups a new baseline for future sentence-level methods developed in a wide set of languages.

Keywords: Sentiment Analysis, Multilingual, Machine Translation, Online Social Networks, Opinion Mining.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Online Social Networks (OSNs) have been used by billions of users worldwide, and it is the most popular Web application nowadays [Shannon Greenwood, 2016]. On those systems, users can discuss an enormous variety of subjects, expressing their opinions, politic views, and even some subjective concepts like sentiments. Because of the massive popularity and quantity of data shared on those systems, a diversity of applications have emerged, aiming at extracting opinions and inferring public sentiments.

In this context, sentiment analysis has become a popular tool for data analysts, especially those that deal with social media data. It is common to find public opinion and reviews of services, events, and brands on social media. From the extracted data, sentiment analysis techniques can infer how people feel about a particular target, which is essential for companies aiming at focusing their investments on incorporating those potential clients and creating a more specific but also massive public marketing. Thus, sentiment analysis became a hot topic in Web applications, with the high demand from industry and academy, motivating the proposal of new methods to deal with this subject. Figure 1.1 gives us the idea of how popular sentiment analysis is nowadays. In the figure, the crescent blue line illustrates the search interest of the term "sentiment analysis" worldwide in Google's engine, given the time period indicated in the x-axis.

Despite the large interest from industry and academy in the sentiment analysis field, substantial effort has been focused on sentiment analysis solutions that depend on the English idiom, since it is dominant across the Web content [Narr et al., 2012]. However, the potential market for sentiment analysis in different languages is vast. For example, suppose a mobile application that simply uses sentiment analysis. To leverage the application to multiple languages and several countries, one would require dealing with sentiment analysis approaches on multiple languages as well, which is currently quite limited. Some efforts even attempt to develop techniques to analyse sentiments from other spe-

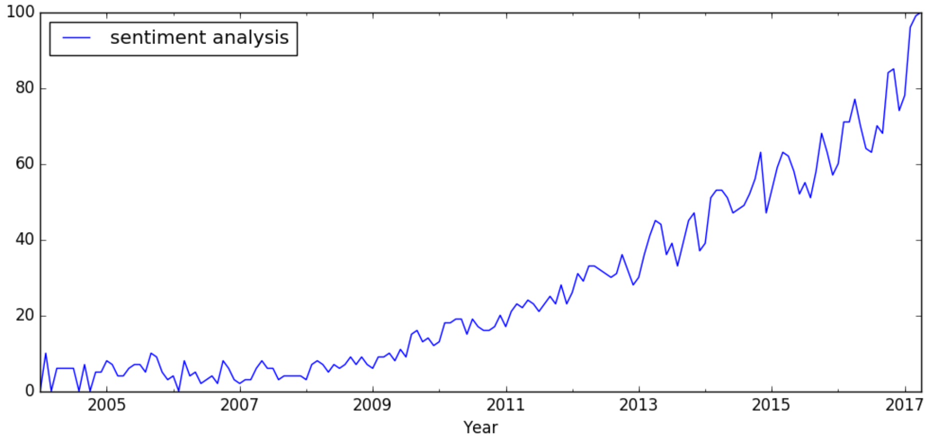


Figure 1.1. Interest in "Sentiment Analysis" since 2004 according to Google Trends

cific languages: Arabic [Abdulla et al., 2013; Refaee and Rieser, 2015], German [Remus et al., 2010], Portuguese [Souza and Vieira, 2012], Russian [Yussupova et al., 2012], Spanish [Shalunts et al., 2016], among others. However, little is known about the performance prediction, viability and real need of those methods. More important, a different solution on each specific language is unfeasible for those interested in using sentiment analysis as part of a system or application that supports multiple languages.

This work investigates how a simple strategy can address the problem of sentiment analysis in multiple languages. Additionally, it arguments towards the use of translation-based techniques as a baseline for new multilingual sentiment analysis methods. Particularly, it analyses how the use of machine translation systems - such as Google Translate[1], Microsoft Translator Text API [2] (used by Bing Translator[3]) and Yandex Translate [4] - combined with state-of-the-art English sentiment analysis methods can be comparable to methods created specifically to non-English texts.

We should emphasize that our work focuses on comparing "off-the-shelf" methods as they are used in practice. [Dashtipour et al., 2016] reproduced many of the methods from the literature and compared their outputs with the reported original results. They concluded that many of the methods are not described accurately, and most of the reproduced results had lower accuracy than previously reported. Therefore, we choose methods that we could reproduce. This excludes most of the supervised methods which require labeled sets for training, as these are usually not available for practitioners. Moreover, most of the supervised solutions do not share the source code or a trained model to be

[1] https://translate.google.com

[2] https://www.microsoft.com/en-us/translator/translatorapi.aspx

[3] https://www.bing.com/translator

[4] https://translate.yandex.com/

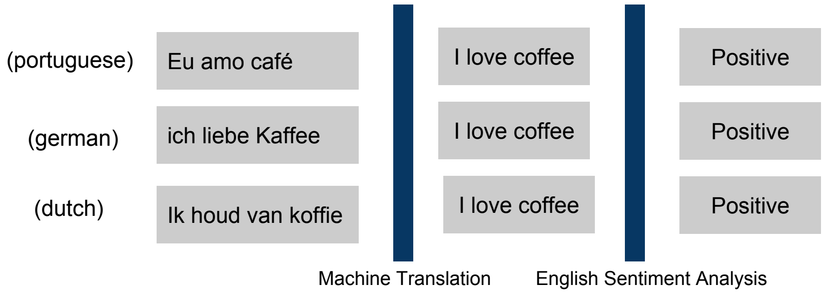


Figure 1.2. Our methodology overview with examples

used with no supervision.

Using the output from machine translation tools, we show an evaluation of the prediction performance of 15 sentiment analysis methods recently evaluated in a benchmark study [Ribeiro et al., 2016] - AFINN, Emoticon Distant Supervisor, Emolex, Emoticons, Happiness Index, NRC Hashtag, OpinionFinder, OpinionLexicon, Panas-t, Pattern.en, SASA, SentiStrength, SO-CAL, Stanford Recursive Deep Model, Umigon, and Vader - across 14 different languages: Chinese, German, Spanish, Greek, Croatian, Hindi, Czech, Dutch, French, Haitian Creole, English, Portuguese, Russian, Italian. According to *Internet World Stats*[5], seven of those languages appear among the top ten languages used on the Web and represent more than 61% of non-English speaker users. In Figure 1.2, we present an overview of the methodology discussed in this work.

Despite the still large existent space for improvement in current state-of-the-art sentiment analysis methods for English, as suggested by a recent benchmark study [Ribeiro et al., 2016], our findings suggest that machine translation systems are mature enough to produce reliable translations to English that can be used for sentence-level sentiment analysis and obtain a competitive prediction performance results. Additionally, we show that some popular language-specific methods do not have a significant advantage over a machine translation approach.

1.1 Objectives

The main objective of this work is to provide a quantitative comparison between the use of several already developed English methods for sentiment analysis combined with machine translation in the multilingual context. Also, we want to compare the results with current language-specific methods in order to identify if these methods are better than the machine translation approach.

[5] `http://www.internetworldstats.com/stats7.htm`

Our hypothesis is based on the assumption that machine translation of datasets to English and it posterior analysis on English specific methods can be as good as the specific sentiment analysis methods created for determined languages. We support this, because, even when words change between two paired sentences in different languages; an accurate machine translation should not change their meaning and its sentiment polarity.

1.2 Results and Contributions

To address the problem of multilingual sentiment analysis we perform several experiments using methods created for English on multilingual datasets with the help of automatic machine translators. As the main result, the hypothesis proposed was confirmed, and the methods designed for specific languages do not overcome in any evaluation the machine translation approach. Although there is not a best sentiment analysis method for all datasets, this work highlights that current commercial and non-commercial methods for sentiment analysis for non-English datasets are not powerful enough against the machine translation approach combined with state-of-the-art sentiment analysis methods published in English.

There are two main contributions for this work. First, our empirics results provide evidence that machine translation, although a simple approach, can be more efficient than "off-the-shelf" methods developed specifically for that language. Therefore, machine translation should be used as a baseline when new methods are proposed by the scientific community. This contribution also motivates scientist who wants to create new sentiment analysis methods for specific languages to investigate niches where machine translation are still not good enough such sarcasm words, slangs and others colloquialisms. As a final contribution, using the concepts discussed in this work, the iFeel system[6] was developed and released for the community as an open and free framework for sentiment analysis in multiple languages.

1.3 Publications

Several publications were made since the begin of this study. Although some of them are not related to multilingual sentiment analysis specifically, all of them contribute in some manner to this work and the sentiment analysis field in general.

On the particular multilingual sentence-level sentiment analysis field, there are two main publications which resulted in this work:

[6] iFeel is hosted on http://www.ifeel.dcc.ufmg.br

- Araújo, M., Reis, J. C., Pereira, A. C., and Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *ACM Symposium on Applied Computing.*

- Reis, J. C., Gonçalves, P., Araújo, M., Pereira, A. C., and Benevenuto, F. (2015b). Uma abordagem multilıngue para análise de sentimentos. In *IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015).*

- Araújo, M., Diniz, J. P., Bastos, L., Soares, E., Ferreira, M., Ribeiro, F., and Benevenuto, F. (2016). ifeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis. In *Tenth International AAAI Conference on Web and Social Media.*

In these publications, we discuss the same approach described here, including a previous version of iFeel. However, now we extended them, adding new language datasets, performing a comparison between machine translators and adding a broader related work section to cover recent approaches.

In the following publications: [Gonçalves et al., 2013], [Araújo et al., 2013], [Messias et al., 2016], [Ribeiro et al., 2016] we compared the sentence-level sentiment analysis methods in several English datasets in order to evaluate their performance like accuracy, and computational resources usage, including on mobile devices. At [Araujo et al., 2016] an emotional signature was proposed to reveal the ability of sentiment analysis to predict a book author.

1.4 Organization

The rest of this document is organized as it follows:

- **Chapter 2 - Sentiment Analysis.** This chapter presents an overview of the main concepts and terminologies related to sentence-level sentiment analysis and the current state-of-the-art methodologies. Furthermore, we describe existing approaches for non-English sentiment analysis including previous direct machine translation approaches as focused in this work and how we distinguish from them.

- **Chapter 3 - Methodology.** This chapter presents the resources used throughout this work in order to evaluate our hypothesis. It describes our effort in gather representative human labeled datasets in multiple languages, the machine translation systems used to translate these datasets to English, and all the English and non-English sentiment analysis methods.

- **Chapter 4 - Experimental Evaluation.** This chapter presents the results and discussions that we use to validate our hypothesis. It describes how we did the performance comparison between machine translation approach and language-specific methods, including an evaluation of machine translation systems and a ranking of best non-English and English methods.

- **Chapter 5 - iFeel System.** This chapter presents a web-based framework for multilingual sentiment analysis named iFeel, developed to facilitate the sentiment analysis study by the community and share the code including the datasets of this work.

- **Chapter 6 - Conclusion.** This chapter presents the findings of this study, it also highlights advantages, disadvantages, limitations and possible improvements related to the future of this work.

- **Appendix A.** Results of Applicability , F1-Score and Macro-F1 for every language datasets and every supported method.

Chapter 2

Sentiment Analysis Overview

2.1 Definitions and Terminologies

Given, the recent popularity of the term sentiment analysis, it has been used to describe a wide variety of tasks by the community. Therefore, there is a broad concept of what sentiment analysis is, and where many subfields have emerged. We can list a range of subfields inside sentiment analysis, for example, detection of polarity in a sentence, evaluation of subjectivity in a text or detection of opinions related to objects of interest. There are a variety of conferences that covers these topics, in particular when related to natural language processing, for example, the Annual Meeting of the Association for Computational Linguistics (ACL) and Conference on Empirical Methods in Natural Language Processing (EMNLP). However, the SemEval[1] workshop highlights as one that annually tries to evaluate the current state-of-the-art techniques and proposes several new challenging tasks for the field. During the current version of the SemEval workshop, five tasks related to sentiment analysis are proposed. Each of these tasks has many subtasks, ranging from three-class polarity detection of tweets to veracity prediction given a rumor.

Since there are many definitions related to the sentiment analysis field, here we list and describe the concepts we use under the context of this work.

- **Polarity Detection:** The detection of polarity in a sentence, is a typical task in sentiment analysis and has many subtasks involved. First, there is a polarity classification problem where the goal is to identify which class of sentiment a sentence have. This task can be a 2-class problem where the method has to determine if a sentence is negative or positive, or it can be a 3-class problem where the neutral class is added. Second, there is a topic-based polarity classification problem, where

[1] http://alt.qcri.org/semeval2017/

given a message and a topic, the goal is to identify whether the message expresses positive, a negative or neutral sentiment regarding that topic. Third, a contextual polarity disambiguation problem where given the word or phrase inside a text, the method should decide which class of sentiment it expresses in that context. Finally, the detection of trends inside a topic, where given some messages during a period of time, the goal is to identify the overall sentiments towards the topic [Rosenthal and Stoyanov, 2015].

- **Strength association with sentiment:** Several sentiment analysis methods rely on labeling a sentence with polarity classes (positive, negative or neutral). However, a recent addition to the polarity classification task is to provide a score related to the detected class. This score represents the intensity of the sentiment in a sentence. Some authors developed methods where the score ranges from -1 to 1, other authors define the range as 0 to 1. Anyway, as higher the score is, more strong is the positive sentiment in the sentence. Correspondingly, as lower the score is, more negative is the sentence [Kiritchenko et al., 2014].

- **Emotion Detection:** In this task, the goal is to indicate an "affective text", or emotion, given a text. Since some words have emotional meaning, like surprise, anger, happiness, the methods should be able to identify correctly the best "affective text" that matches with the sentence. Emotion detection is a hard task due some words have emotions depending on the context, for example, "monster" or "ghost". Some authors predefine a limited list of words that they consider as "affective text". Usually, these lists are from psychology research, as used in [Strapparava and Mihalcea, 2007] and [Goncalves et al., 2013].

- **Multilingual Analysis**: The task of multilingual analysis is a particular case of sentiment analysis where different languages are involved. Some authors consider multilingual analysis when given a text, it can partly be written in one language and partly in some other language [Vilares et al., 2017]. However, the most common use of multilingual sentiment analysis is when authors propose a generic methodology to perform sentiment analysis in datasets written in just one language, usually different than English. The motivation of this problem emerges from the fact that most of the proposed methods for sentiment analysis are targeting only English.

- **Native Method:** This term was coined throughout our recent work [Araújo et al., 2016]. We understand as native methods, the sentiment analysis methods which were built with a certain target language in mind, usually non-English. It can also be a particular version of a model from a method that originally was created for

English, but it had a new training dataset from a different language to perform multilingual analysis. Also, it can be a brand new methodology that investigates particularities of a specific language to build models that can predict the sentiments.

- **Off-the-shelf Methods**: These are convenient methods for sentiment analysis that were published, and the authors share its source code and model for ease use. These methods are used in practice and exclude most of the supervised methods which require labeled sets for training. However, it can include the methods whose models are shared by the authors and are used as an unsupervised method [Ribeiro et al., 2016].

We explored a wide range of tools and methods proposed for this task and observed that they are proposed for different levels of granularity given a document. The granularity level says that the classification given by a method may be attached to whole documents (for document-based sentiment), to individual sentences (for sentence-based sentiment) or specific aspects of entities (for aspect-based sentiment) [Feldman, 2013]. In other words, the lower the granularity, the more specific the sentiment classification is. Next, we better describe these three levels of granularity:

- **Sentence-level**: This granularity level is based on the fact that in a single document there are multiple polarities involved [Pang and Lee, 2008]. This level is often used when we want to have a more fine-grained view of the different opinions expressed in the document about the entities [Feldman, 2013]. Most approaches using this granularity in sentiment analysis are either based on supervised learning [min Kim and Hovy, 2007] or on unsupervised learning [Yu and Hatzivassiloglou, 2003].

- **Aspect-level**: This granularity level is based on the hypothesis that in many cases people talk about entities that have many aspects (attributes) and they have a different opinion about each of the aspects [Feldman, 2013]. In other words, at this level, a sentence can be judged by different entities and may have different polarities associated with it [Pang and Lee, 2008]. This strategy of often used for reviews. For example, the sentence "This hotel, despite the great room, have a terrible customer service" has two different polarities associated with "room" and "customer service" for the same hotel. While "room" has a positive polarity associated with it, "customer service" is judged in a negative way. Many researchers have been using this approach to the sentiment detection task [Popescu and Etzioni, 2005; Wu et al., 2009; Hai et al., 2011]

- **Document-level**: At this granularity level, the polarity classification occurs at the document level, to detect the polarity of a whole text at once. This is considered the simplest form of sentiment analysis and assumes that all the document is related to a single entity, such as a particular product or topic and consequently, associated with a single polarity [Tsytsarau and Palpanas, 2012]. [Pang et al., 2002] show that even in this simple granularity level, good accuracy can be achieved.

In this study, we focus on the use of "off-the-shelf" methods to perform multilingual sentiment analysis. As granularity filter we choose the sentence-level. Moreover, we aim the 2-class polarity classification task (positive, negative) of messages.

Although many of the methods we use, give a strength score associated with the intensity of the sentiment, we map these outputs to the 2-class detection problem. Also, many of the methods have the neutral output. This extra class would transform the problem in a 3-class question as deeply discussed at [Ribeiro et al., 2016]. However, for the purpose of this work, we simplify our experiments, focusing on the 2-class problem.

Next, we discuss a brief introduction to the current state-of-the-art sentiment analysis methods in the case of the English language, and furthermore, we explore the literature related to non-English sentiment analysis techniques.

2.2 English Methods

Due to the enormous interest and applicability, there has been a corresponding increase in the number of proposed sentiment analysis methods in the last years. These methods rely on many different techniques from various computer science fields.

In the case of machine learning, we give as example [Pannala et al., 2016], where the authors discuss the use of Support Vector Machines (SVM's) and Maximum Entropy(EM) regarding polarity detection on aspect-level. Also, [Shi and Li, 2011] proposes a supervised machine learning approach using unigram feature with two types of information (frequency and TF-IDF) to realize polarity classification of hotels reviews. Since there are uncountable papers that explore machine learning, we refer to [Pang et al., 2002] as an interesting learning material and [Cambria et al., 2013] as a resource that discusses most recent sentiment analysis methods which use this technique.

In addition to machine learning, there is the lexical-based approach. In this case, the methods make use of predefined lists of words, in which each word is associated with a specific sentiment. The lexical methods vary according to the context in which they were created. For instance, LIWC [Tausczik and Pennebaker, 2010] was originally proposed to analyze sentiment patterns from formal written English texts, whereas PANAS-t

[Goncalves et al., 2013] and POMS-ex [Bollen et al., 2009] were proposed as psychometric scales adapted to the Web context. Also, [Khan et al., 2015] uses a lexical-based approach to classify sentiment polarities on the aspect-level of granularity.

From the information retrieval field, [Paltoglou and Thelwall, 2010] proposes a sentiment analysis using term weighting functions for TF-IDF technique and compares it in a variety of public datasets reporting significant results. Moreover, [Li et al., 2012] uses active learning approach, named co-selecting, by taking both the imbalanced class distribution issue and uncertainty into account. The authors claim that, with this technique, they were able to reduce the annotation cost for imbalanced sentiment classification.

Overall, the above techniques are acceptable by the research community, and it is common to see concurrent important papers, sometimes published in the same computer science conference, using completely different methods. For example, the famous Facebook experiment [Kramer et al., 2014] which manipulated users feeds to study emotional contagion used LIWC [Tausczik and Pennebaker, 2010]. Concurrently, [Reis et al., 2014, 2015a] used SentiStrength [Thelwall, 2013] to measure the negativeness or positiveness of online news headlines, whereas [Tamersoy et al., 2015] explored VADER's lexicon [Hutto and Gilbert, 2014] to study patterns of smoking and drinking abstinence in social media.

As the state-of-the-art has not been well-defined, researchers tend to accept any popular method as a valid methodology to measure sentiments. In a recent study, [Ribeiro et al., 2016] compared many of the current sentiment analysis methods "off-the-shelf" for the English language over several English datasets, although they claim that exist methods that usually have a better performance than others, they also conclude that there is no best method that can perform the best in all cases. Their study highlights the importance of both main techniques: machine learning and lexicons. Finally, [Gonçalves et al., 2013] and [Gonçalves et al., 2016] also spotlight the potential improvements of performance when combining multiple methods output according to some weighting techniques.

Previously, we described some of the methodologies used in the "state-of-the-art" methods for sentiment analysis in English. Nonetheless, the main focus in this work is multilingual sentiment analysis, therefore, in the next section, we discuss in details many of its solutions. In this case, some of the techniques for English also applies. However, other many strategies emerge. In particular, the ones where English and non-English datasets exchange information.

2.3 Multilingual Sentiment Analysis

Most approaches for sentiment analysis available today were developed only for English, and there are few efforts that explore the problem considering other languages. Besides this disadvantage compared to English, we list in the following subsections several tentatives that move towards a multilingual sentiment analysis context.

In general, these previous efforts focus on adapting strategies that previously succeeded for English to other languages. Overall, they provide limited baseline comparisons and validations. It is unclear if currently available specific language strategies are able to surpass existing sentiment analysis for English if we apply text translation to English.

2.3.1 Machine translation-based methods

This work is not the first to use machine translation as a mechanism to archive sentiment analysis in multiple languages. As an approach similar to our work, [Refaee and Rieser, 2015] performed machine translations in Arabic tweets to English. They show that both strategies, a translation-based and a native method perform equally well. At [Shalunts et al., 2016], the potential of machine translation on sentiment analysis is also explored, using the combination of two state-of-the-art sentiment analysis methods, the authors translate an original corpus from German, Russian and Spanish to English. Then, the results from the translated text are compared with native methods, where, in the worst case it was only 5% inferior. According to the authors, such a setup may be advantageous when lacking the appropriate resources for a particular language and when fast deployment is crucial.

For instance, Banea [Banea et al., 2008] investigates the consequence of automatic corpora generation to sentiment analysis of languages that do not have specific resources or tools. Considering automatic translation to Romanian and Spanish, they investigate the performance of polarity classification from a labeled English corpus.

In another context, [Balahur and Turchi, 2012] investigates the problem of sentiment detection in three different languages: French, German, and Spanish. Their main focus is on evaluating how an automatic translation of text would work to obtain training and test data for these three languages and subsequently extract features that are employed to build machine learning models using Support Vector Machines. Similarly, in order to build a real standalone multilingual sentiment analysis system [Balahur and Turchi, 2013] builds a simple method for English using a Gold Standard dataset and translates this dataset from English to four other languages -Italian, Spanish, French and German to rebuild his sentiment analysis method into a multilingual settings. This work claims that the resultant sentiment analysis can perform multilingual classification with 70% of accuracy.

Nevertheless, our work is the first to test this technique in such wide range covering 14 different languages and comparing the results of 15 English sentiment analysis methods against 3 language-specific methods increasing the confidence in the hypothesis. Besides, all the resources including the iFeel system were developed throughout this work to allow easy access to the methods and techniques by the community, this extra work is unique and helps maintain the reproducibility in the field.

2.3.2 Lexicon and corpus-based methods

On rule-based methods, a set of product features is extracted from a training dataset. These features, or rules, implicates that if the same word from a sentence appears in a previously defined rule, it has a high probability this sentence has the same opinion from the respective rule. From the example given by [Yang and Shih, 2012], considering the following extraction rules: lithium-ion -> battery, mAh -> battery, rechargeable -> battery; they indicate that if "lithium-ion," "mAh," or "rechargeable," appears in a review sentence, we have high confidence to believe that this sentence contains the opinion of a specific reviewer on the product feature "battery" and should be regarded as the opinion of the sentence. Moreover, these rules are built on the combination of lexicons and several linguistic tools such as part-of-speech (POS).

In [Wan, 2008], the authors propose an approach that uses an English dataset to increment the results from a Chinese sentiment analysis. The authors use a set of lexicons to develop a rule-based method which includes: positive and negative lexicons, negation lexicons to reverse the semantic polarity of some terms when convenient, and intensifier lexicons to change the degree of positiveness and negativeness of class. In the same direction, [Abdel-Hady et al., 2014], propose an unsupervised method to analyze polarity in Portuguese and Spanish, based on a language-specific resource (WordNet).

Moreover, [Al-Ayyoub et al., 2015], uses an unsupervised lexicon-based approach to apply sentiment analysis in Arabic. They focus on a 3-class polarity classification problem (positive, negative and neutral) and consider in their solution different components such as various part-of-speech (POS), negations, modifiers, clauses, context. These values were combined with each other to calculate the sentiment value of the sentence. The overall accuracy of this approach was 86%. A disadvantage of this method is that it is not able to handle the many different Arabic dialects.

2.3.3 Machine Learning-based methods

Many of the proposed methods, not limited to this subsection, uses at least in part machine learning techniques. Usually, the most frequent models for classification task are Naive Bayes, Maximum Entropy and Support Vector Machines. While lexical resources are still used to detect the polarity in the text, machine-learning approaches are more common in this type of analysis. Also, machine translation engines are often used in conjunction with various English knowledge bases to generate training data [Lo et al., 2016]. Although these techniques often have a higher performance reported compared to unsupervised approaches, it is also highly depended on the training dataset, inclusively, driven by the context from the source of the collection data.

In [Sidorov et al., 2013], they explore how different settings of text features such as n-gram size, corpus size, the number of sentiment classes can affect the precision of the certain machine learning algorithms for 3-class polarity classification. Also, [Boiy and Moens, 2009], proposes a multi-level/cascading where a subjectivity analysis is done before the detection of polarity. Instead of using a machine translation technique, the authors manually annotated three datasets (English, Dutch and French) to train different machine learning algorithms.

Finally, there other several works that proposes machine learning solutions to different languages including, Arabic [Abdulla et al., 2013], Portuguese [Souza and Vieira, 2012], German [Remus et al., 2010], and Russian [Yussupova et al., 2012].

2.3.4 Parallel corpus-based methods

A different approach to the multilingual solution for sentiment analysis is the use of a parallel corpus that does not depend on machine translation. In this case, the authors acquire some amount of sentiment labeled data and a parallel dataset with the same semantic information, but in different languages. Given the labeled data in each language, the authors exploit an unlabeled parallel corpus based on the assumptions that: two sentences or documents that are parallel should exhibit the same sentiment. Therefore, their goal is to identify and maximize the joint likelihood of the language-specific labeled to infer its sentiment labels [Lu et al., 2011]..

In [Meng et al., 2012], the authors propose a technique named cross-lingual mixture model (CLMM), where they focused on maximizing the likelihood of a bilingual parallel data in order to expand the vocabulary of the target language. The CLMM shows effective when labeled data in the target language is scarce. Also, the authors show that this methodology can boost the machine translated approach where there is a limited vocabulary depending on the machine translator. Their results show an improvement of 12% in the accuracy using this approach when combing corpus from English and Chinese.

A novel methodology using a parallel corpus is also proposed by [Bader et al., 2011]. In this case, the authors use different datasets extract from Bible translations in many different languages. First, they used sentiment-tagged Bible chapters from English to build the sentiment prediction model and the parallel foreign language labels. The authors used others 54 versions of the bible in different languages and the Latent Semantic Indexing (LSI) to converts that multilingual corpus into a multilingual "concept space." In order to prevent a high dependency of the model given the Bible context, a step in their methodology was to shuffle the sentences in each class, a technique that helps break any topic/sentiment association. Their results for accuracy ranges from 72% to 75%.

2.3.5 Hybrid cross-lingual and lexicon-based methods

Many techniques combine corpus-based and lexicon-based approaches, focusing on the domain adaption of sentiment analysis for the resource-poor languages or special domains. These techniques mostly use both annotated corpora and lexicon resources towards learning labels and expand vocabulary. When compared to machine translation of datasets, this technique distinguishes due to the use of transfer learning and not direct association. Also, most of their models are developed using machine learning algorithms.

A recent effort [Lin et al., 2011] proposes a set of seed words (adverbs) that are expanded to train classifiers. The labeled dataset for training in several languages was automatically built considering independent language features, such as *emoticons* [Narr et al., 2012]. They conduct experiments individually and combined analysis for English, German, French, and Portuguese, providing limited evaluations for specific scenarios.

Also [Hiroshi et al., 2004], use machine translation between Japanese and English for parsing and pattern matching on the tree structures that are shared between both languages to create a Japanese sentiment analysis system. They include parsing and pattern matching techniques using a transfer-based machine translation technology to develop a high-precision model. The developed system has a lower learning cost when compared to others approaches.

Since a hybrid approach can combine many of the techniques described, [Ghorbel and Jacot, 2011] uses linguistic features such as POS tagging, chunking, and simple negation forms. In order to improve classification, they extracted word semantic orientation from the lexical resource SentiWordNet. Since SentiWordNet is an English lexicon, they apply a word translation from French movie reviews to English before polarity extraction. Their results were compared to a bag of words baseline.

On the other hand, [Demirtas and Pechenizkiy, 2013] do not archive good results using a cross-lingual framework for analyzing movies and product review datasets in English and Turkish. The authors show that expanding training size with new instances taken from another corpus does not necessarily increase classification accuracy. However, co-training classification with machine translation improved the results when used by semi-supervised learning with unlabeled data coming from the same domain.

2.3.6 Neural Networks-based methods

Neural Networks, or also called deep learning-based methods, recently shows a promising approach for text classification and sentiment analysis [Kim, 2014]. A cascade layers with non-linearities models allows them to build complex functions such as sentiment compositionality, while their ability to process raw signals provides them language and domain independence.

More recently, [Ruder et al., 2016], participated in the SemEval-2016 Task 5. They proposed a convolutional neural network (CNN) for both tasks: aspect extraction and aspect-based sentiment analysis. Their methodology was the top-2 in 7 out of 11 language-domain pairs across other candidates for polarity classification, and top-2 in 5 out of 11 language domain pairs for the aspect-based task. They achieved the best-performing results when analyzing sentiment polarity for English, Spanish, French, and Turkish.

2.3.7 Research Gap

This brief literature overview presents how sentiment analysis is complex, with a variety of tasks and subtasks. Also, authors from many fields try to solve this problem. They bring techniques from psychology, information retrieval, natural language processing and machine learning. Usually, mixing part of them to improve their results. Although sentiment analysis is rich in solutions, it is still centered on the English context. We show many tentatives to perform multilingual sentiment analysis. However, these approaches do not have a successful engagement yet. Most of the current applications are simple and language-specific. Therefore, we understand that a comparison of "off-the-shelf" methods applied to a wide range of languages has significant value for the community. Moreover, we show how a translation-based methodology is promising given its relatively low cost and the high efficiency from current commercial machine translators.

Chapter 3

Methodology

Our methodology to evaluate sentiment analysis in multiple languages involves three key elements. The first is a large set of sentiment analysis methods, designed for English, and commonly used for the same task (i.e. identifying if a sentence is positive or negative). To do that, we performed a large search in the literature and contacted authors to gather a set of the "state-of-the-practice" sentiment analysis methods for English. Section 3.1 describes this effort. Second, we obtain a large set of labeled datasets in different languages to use as the gold standard data. We followed a similar approach of contacting several authors and, in total, we acquired datasets in 14 different languages, described in Section 3.2. As a baseline for comparison we use sentiment analysis systems and tools designed natively to non-English sentences, described in Section 3.3. Finally, we do a brief description of 3 commercial machine translation systems used in this work to perform the translations of the datasets to English.

3.1 English Sentiment Analysis Methods

The term sentiment analysis has been used to describe different tasks and problems. For example, it is common to see sentiment analysis to be used to describe efforts that attempt to extract opinions from reviews [Hu and Liu, 2004], gauge the news polarity [Reis et al., 2015a], as well as for tasks that attempt to measure mood fluctuations [Hannak et al., 2012]. Hence, we restrict our focus on those efforts related to detecting the polarity (i.e., positivity or negativity) of a given text, which can be done with small adaptations on the output of some existing methods, a methodology previously described by [Gonçalves et al., 2013; Araújo et al., 2014].

Our effort to identify a high number of sentiment analysis methods consisted of a systematically search for them in the main conferences in the field and then checking

Table 3.1. Overview of the sentence-level methods available in the literature.

Name	Description	L	ML
Emoticons(Gonçalves et al. [2013])	Messages containing positive/negative emoticons are positive/negative. Messages without emoticons are not classified.	✓	
Opinion Lexicon (Hu and Liu [2004])	Focus on Product Reviews. Builds a Lexicon to predict polarity of product features phrases that are summarized to provide an overall score to that product feature.	✓	
Happiness Index (Dodds and Danforth [2009])	Quantifies happiness levels for large-scale texts as lyrics and blogs. It uses ANEW words (Bradley and Lang [1999]) to rank the documents.	✓	
SO-CAL (Taboada et al. [2011])	Creates a new Lexicon with unigrams (verbs, adverbs, nouns and adjectives) and multi-grams (phrasal verbs and intensifiers) hand ranked with scale +5 (strongly positive) to -5 (strongly negative). Authors also included part of speech processing, negation and intensifiers.	✓	
NRC Hashtag (Mohammad [2012]	Builds a lexicon dictionary using a Distant Supervised Approach. In a nutshell it uses known hashtags (i.e #joy, #happy etc) to "classify" the tweet. Afterwards, it verifies frequency each specific n-gram occurs in a emotion and calculates its Strong of Associaton with that emotion.	✓	
SASA (Wang et al. [2012])	Detects public sentiments on Twitter during the 2012 U.S. presidential election. It is based on the statistical model obtained from the classifier Naïve Bayes on unigram features. It also explores emoticons and exclamations.		✓
PANAS-t (Goncalves et al. [2013])	Detects mood fluctuations of users on Twitter. The method consists of an adapted version (PANAS) Positive Affect Negative Affect Scale Watson and Clark [1985], well-known method in psychology with a large set of words, each of them associated with one from eleven moods such as surprise, fear, guilt, etc .	✓	
EmoLex (Mohammad and Turney [2013])	Builds a general sentiment Lexicon crowdsourcing supported. Each entry lists the association of a token with 8 basic sentiments: joy, sadness, anger, etc defined by Plutchik [1980]. Proposed Lexicon includes unigrams and bigrams from Macquarie Thesaurus and also words from GI and Wordnet.	✓	
SentiStrength (Thelwall [2013])	Builds a lexicon dictionary annotated by humans and improved with the use of Machine Learning.	✓	✓
Stanford Recursive Deep Model (Socher et al. [2013])	Proposes a model called Recursive Neural Tensor Network (RNTN) that processes all sentences dealing with their structures and compute the interactions between them. This approach is interesting since RNTN take into account the order of words in a sentence, which is ignored in most of methods.	✓	✓
Umigon (Levallois [2013])	Disambiguates tweets using lexicon with heuristics to detect negations plus elongated words and hashtags evaluation.	✓	
VADER (Hutto and Gilbert [2014])	It is a human-validated sentiment analysis method developed for twitter and social media contexts. VADER was created from a generalizable, valence-based, human-curated gold standard sentiment lexicon.	✓	
Google Prediction API (Google [2017])	The Google Prediction API is a generic machine learning service which has an trained model for sentiment analysis in English out-of-the-box. The API allows you to train your own model, but it is not our goal in this work. It is the only paid method we used to analyse English sentences.		✓

their citations and those papers that cited them. It is important to notice that some methods are available for download on the Web, others were kindly shared by their authors under request, and a small part of them was reproduced from a paper that describes the method. This usually happened when authors shared only the lexical dictionaries they created, letting the implementation of the method that uses the lexical resource to ourselves. Table 3.1 presents an overview of the methods used in this work, the reference paper in which they were published and the main technique they are based on (machine learning or lexicon). As summarized in Table 3.2, we slightly modified some methods to adequate their output formats to the polarity detection task where the output is -1 (negative), 0 (neutral) or 1 (positive). The original output of these methods are written in the table, but we colored as blue the outputs we consider as positive, red the negative output and black what we considered as neutral. The methods used in this work were deeply discussed and had their performance compared throughout different English datasets at [Ribeiro et al., 2016]. Following their methodology we choose 15 methods from that study.

Finally, we also choose to add the Google Prediction API, a commercial sentiment analysis tool created by Google in order to verify the results discrepancies between paid and unpaid methods. All of the methods, excluding the Google Prediction API can be used on the iFeel system developed in this work and described on the Chapter 5.

Table 3.2. Overview of the sentence-level methods

Methods	**Original Output**
AFINN	**-1, 0, 1**
Emoticons	**-1, 1**
Opinion Lexicon	**-1, 0, 1**
Happinnes Index	**1, 2, 3, 4, 5, 6, 7, 8, 9**
SO-CAL	**[<0), 0, (>0]**
NRC Hashtag	**sadness, anger, fear, disgust, anticipation, surprise, joy, trust**
MPQA	**Negative, Neutral, Positive**
Emolex	**negative, positive**
Umigon	**Negative, Neutral, Positive**
Vader	**-1, 0, 1**
PANAS-t	**fear, sadness, guilt, hostility, shyness, fatigue, attentiveness, joviality, assurance, serenity, surprise**
SASA	**Negative, Neutral, Unsure, Positive**
Stanford	**very negative, negative, neutral, positive, very positive**
SentiStrength	**-1, 0, 1**
Google Prediction API	**-1.., 0, ..1**

3.2 Human Labeled Datasets

In this section, we present an overview of the datasets used in this work to compare the performance of our approach against traditional methods. These workloads consist of 14 gold standard datasets of sentences, which were labeled by humans as positive, negative or neutral according to their sentiment polarity. Using the human labels, we can compare the quality of the sentiment analysis methods and judge their performance. In Table 3.3 we summarize the relevant information about these datasets, showing in each row the language, its ISO 639-1 two letter code, the place it was first published, the type of data collect, and the number of positive (Pos) and negative sentences(Neg) labeled by humans[1].

It is important to highlight the process of acquiring these datasets. We contact many others who published works related to non-English sentence-level sentiment analysis, the

[1] The datasets used in this paper are available under request.

Table 3.3. Gold standard labeled datasets

Language	Neg	Pos	Published at	Code	subtype
Chinese	432	446	Wan [2008]	zh	product reviews
German	239	353	Narr et al. [2012]	de	tweets
Spanish	350	683	Villena Román et al. [2013]	es	tweets
Greek	3189	2131	Makrynioti and Vassalos [2015]	el	tweets
French	321	341	Narr et al. [2012]	fr	tweets
English	998	1595	Narr et al. [2012]	en	tweets
Croatian	467	1658	Glavaš et al. [2013]	hr	food reviews
Hindi	230	340	Arora [2013]	hi	product review
Dutch	43	77	Tromp [2012]	nl	tweets
Czech	2808	1422	Kincl et al. [2013]	cs	movie reviews
Haitian Creole	734	128	Ríos et al. [2014]	ht	tweets
Portuguese	414	626	Narr et al. [2012]	pt	tweets
Russian	416	333	Koltsova et al. [2016]	ru	tweets
Italian	1422	820	Basile and Novielli [2014]	it	product reviews

result of this extensive manual work is a unique and rich source of human labeled sentences in many languages. It is very challenging to produce datasets labeled by human regarding sentiments because of two main reasons: the subjectivity intrinsic in the sentence (context dependent) and the amount of time needed to humans label thousands of sentences. In our case we would have an extra laborious task due the multilingual context, considering that humans who label these datasets should know fluently these many different languages. So, the manner we found to successfully proceed with this work was contacting different and independent authors in the field who already did this labeling work in a specific language. After getting these 14 independent datasets, we post-process them to make sure the labels are all the same and can be comparable to the sentiment analysis methods output.

Notice that not all of the datasets took the same research policies and standards. Some were labeled by three humans others by two humans. Some used Amazon Mechanical Turks and others used what the authors called specialists. Also, some of them were collected with a theme in mind, for example, the author of the Russian dataset collect data about product reviews in Russian blogs, differently, the Croatian dataset author focused on food reviews, others are formed by random tweets. We perceive these differences in the proceedings to generate the gold labeled datasets as the biggest limitation of our work since we are comparing sentences from different sources and contexts labeled by different policies according to each researcher. However, we also understand the goal of sentence-level sentiment analysis is be generic and independent of the context. Thus, we treat all the datasets equally without configuring or training the methods for a specific situation.

Since our main goal is to support the hypothesis that native methods do not perform as well as translated datasets, this approach works for our needs.

3.3 Language-Specific Sentiment Analysis Methods (Native Methods)

Ideally, we would like to compare the use of machine translation using all the methods designed for English described in Section 3.1 with a large number of methods proposed for some specific language. We contacted authors of some identified efforts asking for datasets and their methods. While we succeeded in obtaining a large number of datasets, most of these methods are not available even under request to authors, making reproducibility almost impossible in most of the cases. Therefore, we choose to use "off-the-shelf" methods as baseline.

We were able to assess 3 native methods created or trained specifically for certain languages. In Table 3.5 we list and describe these methods shortly and in Table 3.4 we show the list of languages supported by them.

First, we have the Multilanguage version of Sentistrength (ML-Sentistrength), available from the same authors of the original Sentistrength version. These authors released an adaptation of the original sentistrength that consists in changing the lexicons files for the correspondent ones of the language you desire to perform sentiment analysis. In their website, there are available 9 set of lexicons for different languages. This version is free for scientific purpose.

Second, we use a commercial sentiment analysis API namely Semantria [2], which provides results in 21 languages. We used the trial version of the Microsoft Excel Plugin available on their website.

The third baseline is the IBM Watson API, a commercial sentiment analysis toolkit developed by IBM, which has a range of features such as polarity detection, pos-tagging, and others cognitive systems available. For the sentiment analysis purpose, IBM Watson is able to classify the polarity of the sentences in 9 languages.

Notice in Table 3.4 that 4 languages do not have any native sentiment analysis method to compare with, they are Croatian, Hindi, Czech and Haitian Creole. Although the comparison results in this work refers to the languages that have at least one native method that supports it, we still understand that is relevant to show the results for all languages that we have access to human labeled datasets. After all, we can still compare

[2] `https://semantria.com`

the performance between English methods and show the baseline for future authors who wants to create native methods.

Table 3.4. Support Languages Table

Language	Semantria	IBM Watson	Sentistrength
Chinese [3]	✓		
Russian	✓	✓	
German	✓	✓	✓
Spanish	✓	✓	
Greek			✓
French	✓	✓	✓
Italian	✓		✓
Croatian			
Hindi			
Czech			
Dutch	✓		
Haitian Creole			
Portuguese	✓		✓

Table 3.5. Overview of the sentence-level native methods.

Name	Description	Paid
Semantria Lexalytics [2017]	It is a paid tool that employs multi-level analysis of sentences. Basically it has four levels: part of speech, assignment of previous scores from dictionaries, application of intensifiers and finally machine learning techniques to delivery a final weight to the sentence. It was aquited by the Lexalytics in 2013 and listed by GigaOm as one of the top deep learning startups Wikipedia [2017]	✓
IBM Watson API (Alchemy API) IBM [2017]	It is an hybrid approach which incorpores both linguistic and statistical analysis techniques to lead into a single unified system with high accuracy. The system does not only polarity analysis but also document-level, entity-level, keyword level, directional-level and relational sentiment analysis.	✓
ML-Sentistrength Thelwall [2013]	This is a modified version of the original Sentistrength method created for English. The authors released trained lexicons files that subistites the English version in order to support 9 extra different languages. This is multilanguage version is free for scientific purpose	

3.4 Machine Translation Systems

Since 1950s, machine translation or automated translation is a field for research [V. Le and Schuster, 2016]. Its main goal is provide text translation by a computer without human interaction. There are three main approaches to solving the problem of automatic

[3] Simplified/Standardized Chinese

translation: Rules-based/phrase-based, statistical methods or neutral networks. Rules-based uses lexicons combined with grammar definitions in order to translate sentences in a meaningful way. The statistical system tries to build a translation model to analyze a large amount of training data for each language correspondence. Neutral Network based systems build one large neutral network with an huge amount of training data, this approach has recently become popular and shows better translation performance. For the purpose of this work, we want to justify two main potential questions related to the use of machine translators: Why we choose commercial machine translators tools instead of free published tools? And, why we believe that machine translated texts to English combined with English sentiment analysis tools are better than native non-English sentiment analysis methods?

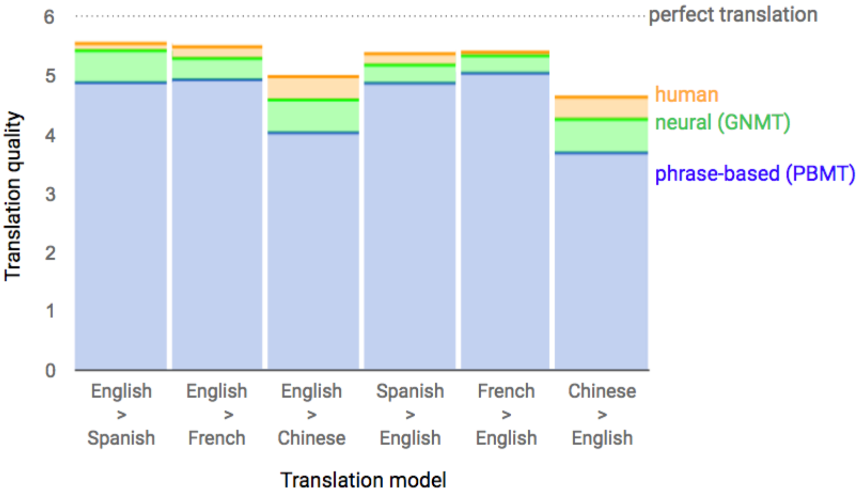


Figure 3.1. Comparison between phrase-based and neural network techniques with a human baseline, extracted from [V. Le and Schuster, 2016]

To answer the first question, we need to clarify that there are available many free open sources machine translation tools for multiple languages[4]. However, these tools even when based on a pre-trained statistical system are static and do not follow the evolution language of the Web. In other words, in such dynamic environment as the Internet, new emoticons, slangs or even ways to express are frequently generated, requiring constant training models [McKelvey, 2016]. So, we choose well-known commercial tools which retrain periodically their models, as explained by [Microsoft, 2017], [Yandex, 2017], [V. Le and Schuster, 2016]. Since we don't have either resources or knowledge to keep an updated trained model of high accuracy in our environment and it's not the purpose of this work do so, we decided to use the commercial API's.

[4] http://fosmt.org/

In Figure 3.1 we see a comparison performance between three translators candidates, a Neural Networks, a phrases-based system, and proper humans. It illustrates how close the current state-of-the-art machine translation systems are to humans translators. Also, it shows that the neural networks approach seems to overcome the phrased-based strategy. So, we answer the second question of the first paragraph combining these results with the axiom that, words will potentially change between two paired sentences in different languages, however, an accurate machine translation will not change their sentiment polarity.

In our work, we used 3 commercial translation tools to translate our non-English datasets to English, they are listed in Table 3.6. The Yandex API allows the user send 10,000 free requests per month. Otherwise, Google Translator has a free Web interface but no free API support, however, when you create an account in the Google Cloud Platform you are granted with U$300,00 to use throughout one year, with a cost of U$20 per million of characters translated. The Microsoft Translator Text API can be used though Microsoft Azure platform, and it allows to process the first 2 million characters for free and for each additional million of characters it costs U$10. Similarly to Google Cloud, the Azure platform also gives $200 dollars to start using their service.

Table 3.6. Overview of Machine Translators tools used

Translator	Description
Yandex Translate API [Yandex, 2017]	Yandex machine translation is based on the statistical approach. To learn a language, the system compares hundreds of thousands of parallel texts that translate each other "sentence by sentence." It has two main components: the translation model and the rule-based model.
Google Translator [V. Le and Schuster, 2016]	Previous version of Google Translator used to be phrase-based and uses English as an intermediary language to translation. However, now it utilizes Neural Networks and direct language paired translation, according to authors this new approach is responsible for improving system performance by 55% compared to phrase-based version.
Microsoft Translator Text API [Microsoft, 2017]	Since 2010 Microsoft uses Neural network in their translation systems. Given any language pair to translate, the system uses unique characteristics from the pair which presents a 500-dimension vector. It encodes concepts like gender (feminine, masculine, neutral), politeness level (slang, casual, written, formal, etc.), type of word (verb, noun, etc.) and other non-obvious characteristics

Chapter 4

Experimental Evaluation

In this chapter, we present all the experiments performed in this work to sustain our hypothesis. We believe that current sentiment analysis methods create for English combined with the current state-of-the-art machine translation system are able to be as good as or even better than native sentiment analysis methods in multiple languages. Several experiments were performed in order to evaluate the following questions: (i) How choosing a translation platform impacts in the overall performance? (ii) What are the performance of English methods for sentiment analysis for non-English content with the help of automatic machine translators? (iii) Is machine translated approach better than native methods? (iv) Is there a difference when considering the performance of these methods only for the analyze of positive or only for negative polarities (v) In which cases, the native methods are better than machine translation approach?

We present in the following section the metrics we choose to analysis the performance of the English and non-English sentiment analysis methods. After, we show the analysis and experiments throughout the chapter and a final summary of the results.

4.1 Metrics

The **F1-Score** is a metric used to compare the quality of the prediction for a given ground truth. In our case, we use it to check how a method is able to identify a sentiment in a sentence related to human labels. The F1-Score considers equally important the correct classification of each sentence, independently of the class, and basically measures the capability of the method to predict the correct output. This metric can be easily computed for 2-class experiments using the Table 4.1.

In this case, the precision of positive class is computed as:

		Predicted	
		Positive	Negative
	Positive	a	b
Actual	Negative	c	d

Table 4.1. Confusion Matrix

$$P(pos) = \frac{a}{(a+c)}$$

The recall is calculated as:

$$R(pos) = \frac{a}{(a+b)}$$

So, the F1-Score for the positive class is:

$$F1(pos) = \frac{2P(pos) \cdot R(pos)}{P(pos)+R(pos)}$$

A variation of the F1-Score is namely **Macro-F1** , it is normally reported to evaluate classification effectiveness on skewed datasets, when the class distribution is not homogeneous. Hence, it is the one we use during our analysis. Macro-F1 values are computed by first calculating F1 values for each class in isolation, and then averaging over all classes. This metric considers equally important the effectiveness in *each class*, independently of the relative size of the class. In our analysis, we only considered the sentences where the method could indicate one of the 2-class, negative or positive, to compute the Macro-F1 . Therefore, the Macro-F1 reported represents how effective the method is when it indicates a polarity.

Although we only use the output of methods that indicates a polarity to calculate the Macro-F1 , the methods still have the neutral classification for some of the sentences. So, we define as **Applicability** , a metric to determine the percentage of sentences a method can, in fact, classify as positive or negative (not neutral). This is important in our work since all the human labeled datasets are fully classified as positive or negative, many of the sentences do not receive any score by the methods. Moreover, it seems that methods which are conservatives regarding given a polarity to sentence usually have a higher accuracy. For instance, suppose that Emoticons' method can classify only 10% of the sentences in a dataset, corresponding to the actual percentage of sentences with emoticons. It means that the Applicability of this method in this specific dataset is 0.1. Note that, the Applicability is quite an important metric for a complete evaluation in the 2-class experiments. Even though Emoticons presents high accuracy for the classified phrases, it was not able to make a prediction for 90% of the sentences. More formally, Applicability is calculated as the number of total sentences minus the number of undefined sentences, all of this divided by the total of sentences.

Throughout the analysis of our results, we mainly discuss the results and tradeoff between these two metrics: Macro-F1 and Applicability . We could propose a new metric based on the product of both. However, we understand that the Macro-F1 might not have the same weight of Applicability depending on the task, hence, during our analysis, we will show and discuss these metrics separately.

4.2 Comparison Between Machine Translators

In this section, we evaluate if there is a difference in the outcome results by choosing the machine translators system to perform our approach. So, using the 3 machine translators we selected to test our hypothesis, all the language datasets were translated from their original texts to English. An exception is the English dataset used only to have a comparison baseline.

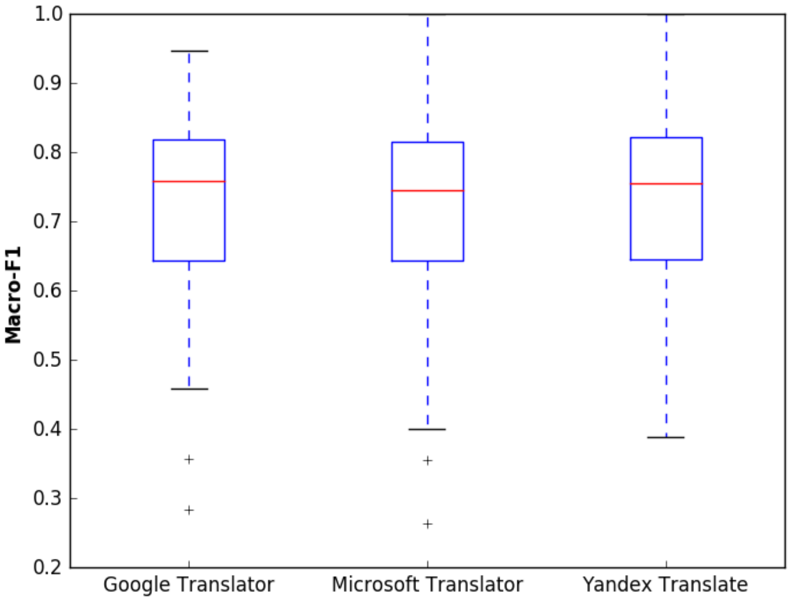


Figure 4.1. Macro-F1 distribution given machine translation system

In Figure 4.1, we present the performance distribution in a boxplot, with the result for all datasets given a particular machine translation system. The distribution is very similar, especially between 25th and 75th percentile, with Google Translator slightly better than others. According to our results, when averaging the Macro-F1 for all methods in all datasets, the systems from Yandex and Google have scores 0.72 with a standard deviation of 0.12. The Microsoft Translator has a marginally inferior performance with an Macro-F1 average of 0.69 and standard deviation of 0.20. Despite this difference, the confidence intervals of the results overlap for $\alpha = 0.95$ and the variation coefficient is 0.02. Therefore, when considering the polarity of the sentences when they are automatic translated, all 3 seems to have consistency and do not change the polarity by their own. In fact, we check all the 472k outputs for all English methods running on translated text throughout this work, and we see that in 3k cases the methods had at least one negative results and one positive result for the same sentence. This means that only in 0,6% of the output sentences from the machine translators have inverted polarities. This conclusion doesn't mean that the sentences are keeping their sentiment polarity from the original language, but it gives confidence that choosing the machine translation system might not impact abruptly in the results.

It's important to explain why the boxplot has so large tail with Macro-F1 outlines close to 0 and 1. These are the case when methods such as Emoticons or Panas-t have poor Applicability . Thus, their Macro-F1 are calculated based on a small sample with high variance. In further sections, we also show the results for both metrics, which helps to understand this tradeoff.

From now on, all the Macro-F1 and Applicability scores discussed in this work are the majoring votings between these 3 machine translators. For example, if the method SOCAL predict the polarity of a sentence as 1(positive) when translated using Microsoft Text Translator and Google translators, but gives -1 (negative) to Yandex translation, we say that SOCAL is positive for this specific sentence.

4.3 Overall Performance

In this section, we go deep into a detailed experimental evaluation of the results we generated. First, we present Figure 4.2 on which is plotted the distribution of Macro-F1 scores for non-Native methods on each language dataset. To complement this Figure, we have at Appendix A, Table A.1 to Table A.14 where we show the results for Applicability , F1-scores (positive and negative classes), and Macro-F1 for each language dataset. Additionally, we have Figures 4.3, 4.4 and 4.5 where we can visualize the behavior of the methods regarding Applicability and Macro-F1 simultaneously. Now, we discuss the main findings regarding these results.

We show the overall performance for each language in the Figure 4.2, here, we want to share one thinking. If you remove the labels on the x-axis is very hard to tell accurately which bar corresponds to the English language. This characteristic indicates that,

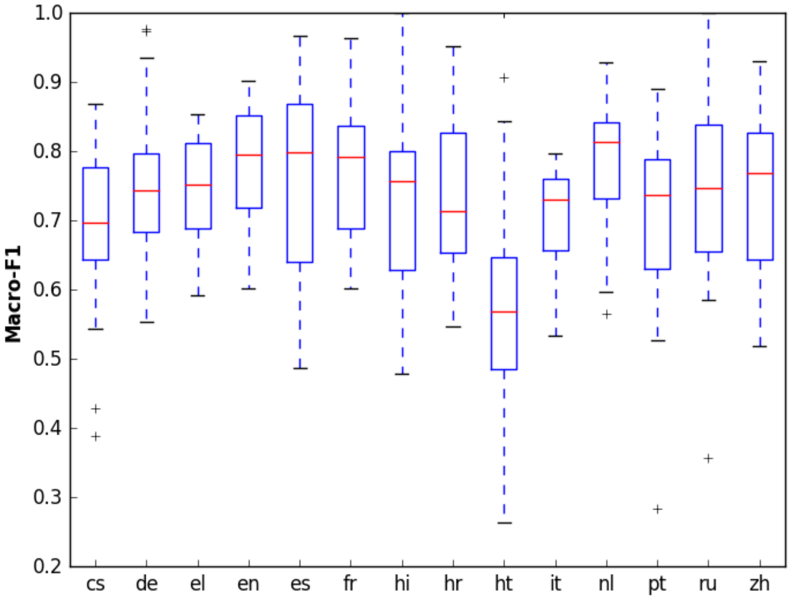


Figure 4.2. Overall performance using our approach on multilanguage datasets

although the datasets were created under different circumstances as discussed in Chapter 3, a potential lack of efficiency of the machine translation approach does not seem to influence the overall results. If the contrary happens, we would expect the corresponding English boxplot as an outline. The only visual outline is the performance of the Creole Haitian dataset, which has a Macro-F1 average below 0.6. Since the Creole Haitian is a language not widely spoken outside Haiti, it has a lack of parallel training data for machine translators; this fact might be the cause of the poor performance observed [Lewis, 2010]. Although this plot gives us an interesting overview of the performance of the methods, especially compared to the English dataset, a deep investigation is needed to fully understand the performance of these methods. So, next, we look into the separated results for each method in each dataset, considering the Macro-F1 and Applicability .

In Figure 4.3, 4.4 and 4.5 we can visualise the tradeoff between Macro-F1 and Applicability previously emphasized. In these figures, we plot the position of each method in a chart, for every language dataset, according to its Applicability (x-axis) and Macro-F1 (Y axis). Therefore, as more close to the upper-right corner of the chart, better the method is. We also highlight the native methods, giving them a red circle. If a method is not shown in the chart, it's Macro-F1 is 0 or it does not support the language.

In these charts, we can see that Emoticons(2), usually appear in the upper-left positions, demonstrating its good Macro-F1 and poor Applicability . Also, we show SO-CAL(13), Stanford(12) as the best method for Chinese with both Macro-F1 and Applicability above 0.8. In the Portuguese dataset, only VADER(11) and Emoticons have Macro-F1 above 0.8, but Vader has a much better Applicability . Finally, Google Sentiment Analysis API(14) highlights as a good approach, actually, this method has a very high Applicability appearing on the right side of most all the charts. Also, its Macro-F1 is often above 0.8. As discussed before, the Haitian Creole chart has the most heterogeneous shape, with many of the methods towards the bottom-left corner.

For instance, regarding the performance of the native methods, we can highlight the IBM Watson(16) for English in Figure 4.4, with an outstanding performance in Applicability and Macro-F1 , on the other hand, it is in the bottom-left corner for French. Further, the Semantria(15) appears with good performance for Chinese, Dutch, Spanish, English, and German, in which it has a Macro-F1 above 0.8, but in several datasets, its Applicability is below 0.5. The Sentistrengh Multilingual (17) appears in these charts with a modest performance, always ranging between 0.6 and 0.8 for both Applicability and Macro-F1 .

The visual findings that we can observe in Figures 4.3, 4.4 and 4.5 also manifests in the data presented at Tables from A.1 to Table A.14 . In these tables, we can identify a strong variation accurately on the prediction performances of some methods for each different language. For example, the Emoticons obtained a Macro-F1 of 1 for the translated Russian dataset, which is much better than the 0.52 obtained for the Spanish dataset. However, it considers most of the instances as neutral (98%) due to the lack of emoticons. This emoticon dependency leads the method to a bad performance regarding Applicability for most of the datasets.

Since these tables show the F1-Score per classes, we can analyze the performance of the methods separately and understand if one is better for analyze positive than negative sentences, or vice versa. For example, several methods have very good performance for one class and a contradictory performance in another. This is the case of the Watson IBM analyzing French where it could evaluate well negative sentences ($\mathit{F1\text{-}Score}_{neg} = 0.86$), but it did not evaluate any of the positive sentences correctly. However, when considering the German, IBM Watson performed much better with the right balance between F1-Score for each class and Applicability . For Croatian, the Happiness Index showed very well in positive sentences. Otherwise, it was poor in predict negative sentences correctly. We noticed by the analysis of the $F1 - score$ per class that most methods are more accurate in correctly classifying positive than negative text, suggesting that methods can lead to bias in their analysis towards positivity.

Still considering Table A.1 to Table A.14 , we notice that some methods obtain consistent results for Macro-F1 still keeping high values of Applicability across multiple languages, such as SentiStrengh, Umigon, SO-CAL, Vader, and Google Sentiment Analysis API. This suggests that these methods might be the most reliable ones for sentiment analysis based on machine translation in the languages analyzed.

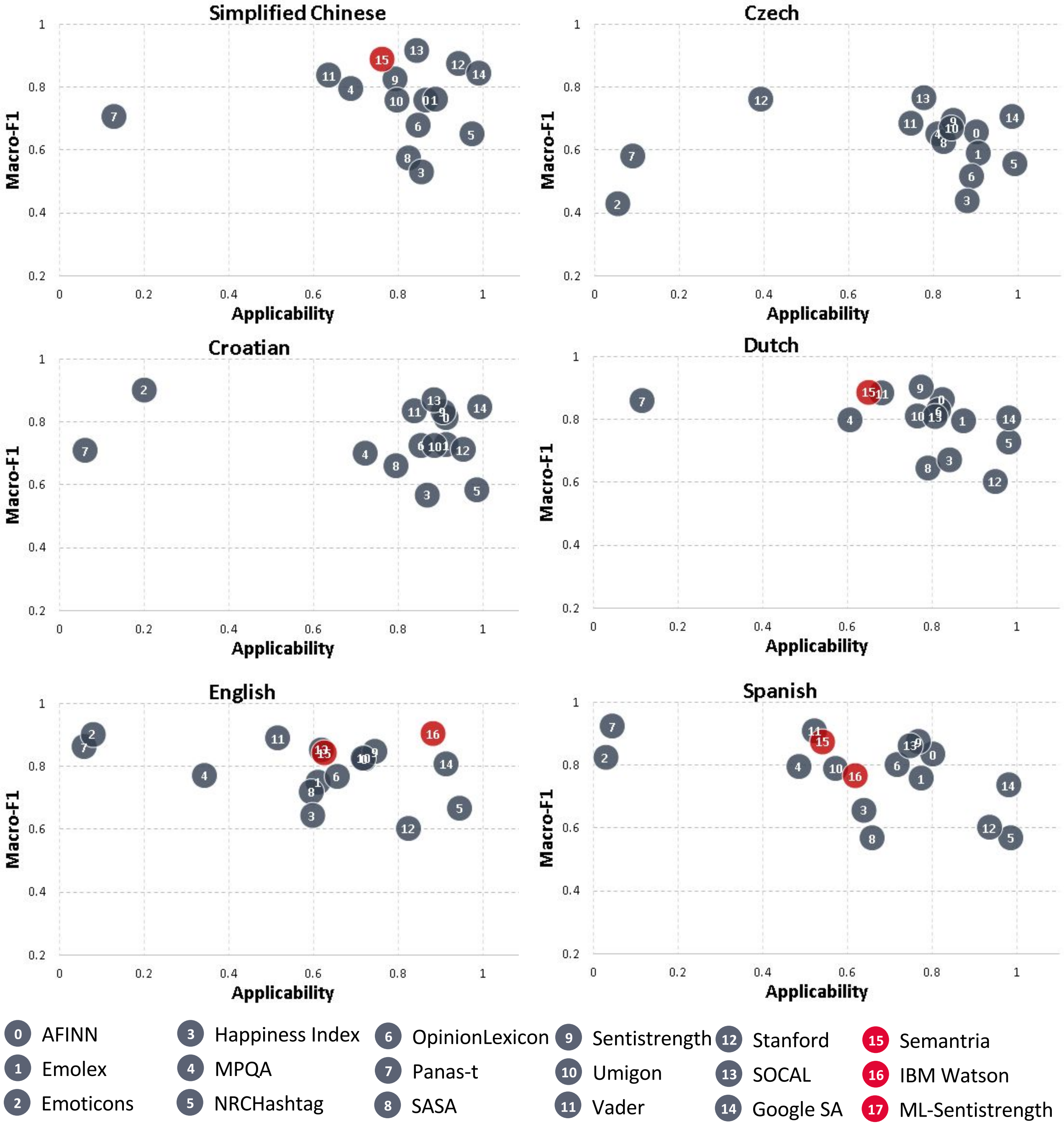


Figure 4.3. Macro-F1 vs Applicability

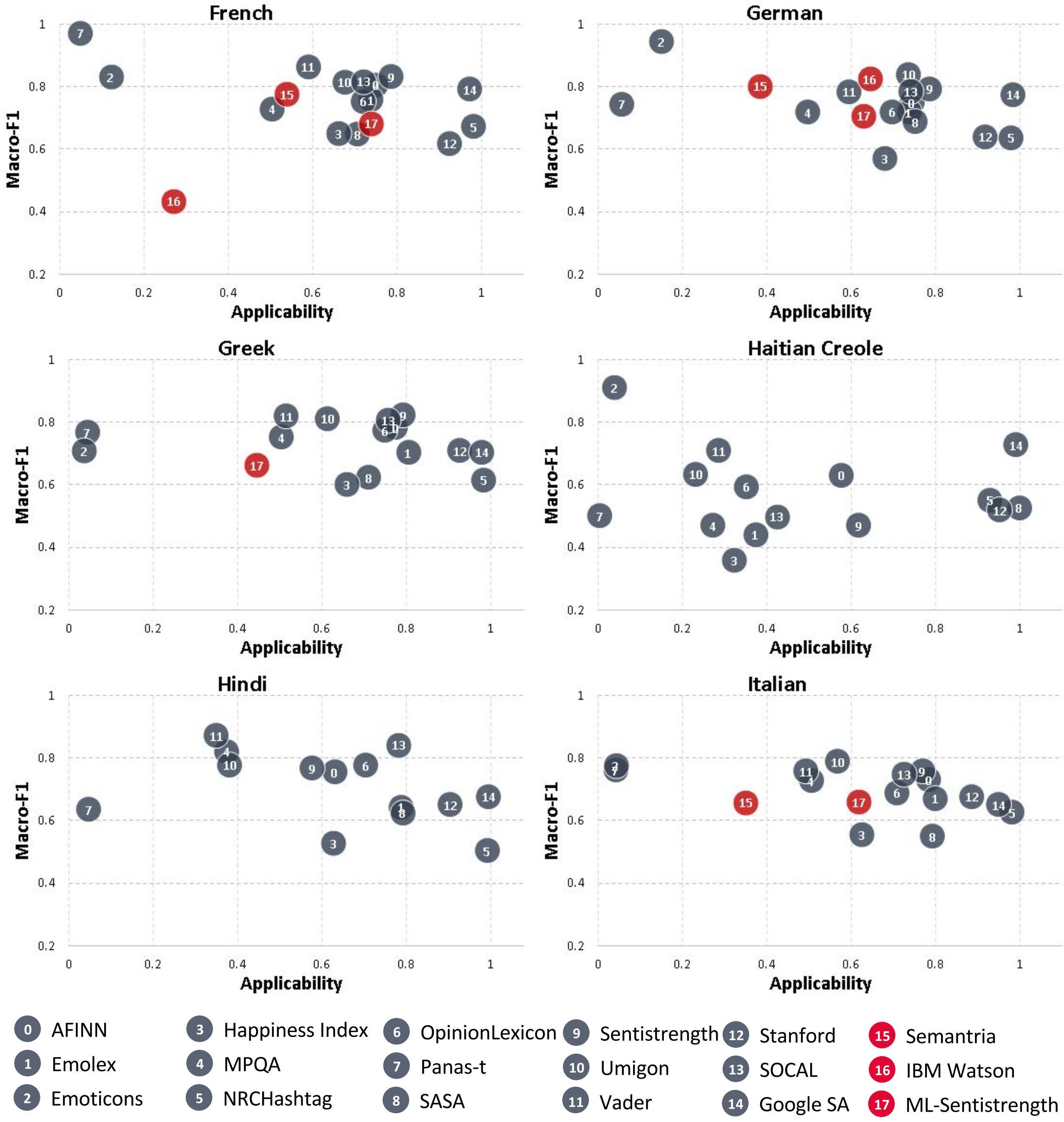


Figure 4.4. Macro-F1 vs Applicability

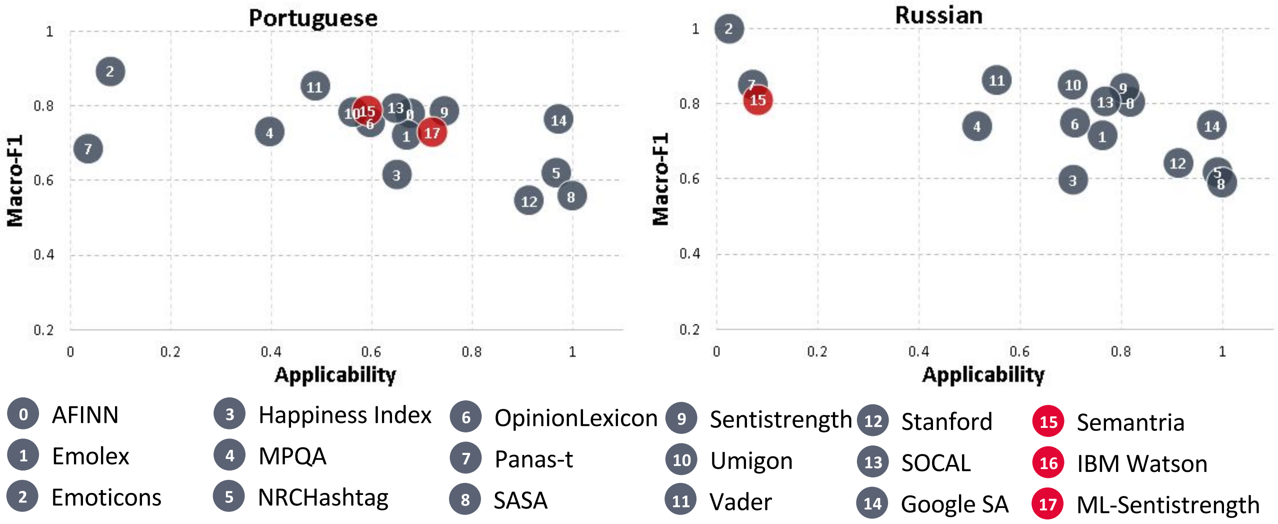


Figure 4.5. Macro-F1 vs Applicability

Table 4.2. Mean Macro-F1 and Applicability metrics comparing the machine translation approach and native methods

Language	Macro-F1	Applicability	Macro-F1 - Natives	Applicability - Natives
Simplified Chinese	0.70	0.74	0.89	0.76
German	0.74	0.67	0.78	0.56
Spanish	0.77	0.65	0.82	0.58
Greek	0.73	0.66	0.66	0.45
French	0.77	0.66	0.63	0.52
Croatian	0.75	0.79	-	-
Hindi	0.66	0.60	-	-
Dutch	0.73	0.72	0.89	0.65
Czech	0.62	0.73	-	-
Haitian Creole	0.57	0.49	-	-
English	0.78	0.60	0.87	0.76
Portuguese	0.72	0.63	0.76	0.66
Russian	0.76	0.69	0.81	0.08
Italian	0.70	0.65	-	0.48
Mean	0.71	0.66	0.79	0.55

Since our main point is to evaluate if machine translation-based methods are able to perform sentiment analysis as well as the natives methods. We summarize the results, separating both groups of methods. In Table 4.2 we present the average for Macro-F1 and Applicability for each language dataset and a final average performance for each group of methods. We can observe that native methods have a higher Macro-F1 score in average, but a lower Applicability . However, some details of these are important to discuss.

In the Russian dataset, for example, the high Macro-F1 for natives come with the cost of only 0.08 in Applicability . Also, the main problem with this evaluation is that we are considering 15 translation-based methods, many of them, push down the Macro-F1 average for the whole group. Therefore, we want to check if there is a subgroup of these methods where we can constantly affirm that they are better than the native methods. In the next section, we provide a different perspective of our results presenting the methods according to the average rank in each dataset. This approach allows us to conclude some interesting findings of our research.

4.4 Ranking the methods

In the previous section, we presented the detailed results generated in this work comparing the Macro-F1 and Applicability metrics between machine translation approach and native methods. Moreover, we grouped the results from each approach in order to compare both techniques. Although the results indicate that machine translation can outperform natives methods, it is not clear which methods should we choose to perform the multilingual analysis. Now, we will present another perspective of our results showing a rankings of the methods based on the average position of them in each dataset.

To build these ranks, we considered separately each metric(Macro-F1 or Applicability). First, for each language dataset, we rank all methods according to one of the metrics. Then, we summarize our results in a table, where, in one column we show the average position of each method with its confidence interval ($\alpha = 0.95$) give the rankings across datasets, and in another column, we show the average score of the chosen metric. So, in Table 4.3 we show these results considering the Macro-F1 , and in Table 4.5 we show the results for Applicability .

In Table 4.3, the methods Emoticons, Vader, SOCAL and Sentistrength are shown as the best methods to analyze these datasets. Semantria has a relatively good Macro-F1 average compared with them, where is only 0.01 behind Emoticons and Vader, but its average position appears at 5th in the rankings. After Semantria, the best native method is the IBM Watson, but with a Macro-F1 average of 0.67. Thus, according to our results and when evaluating only the average position in the rankings based on Macro-F1 , we conclude that machine translation approach seems to be better, and can be comparable to native methods. Therefore, next, we evaluate the average position performance based on the Applicability metric.

Now, considering Table 4.5 where Applicability is taking into account, we have interesting findings. First, the Google Sentiment Analysis API and NRCHashtag appears in the

Table 4.3. Average ranking position considering **Macro-F1**

Method Name	**Average Ranking**	**Mean Macro-F1**
Emoticons	1.50 (±1.19)	0.87
Vader	2.71 (±0.95)	0.83
Sentistrength	4.07 (±1.24)	0.80
SOCAL	4.29 (±1.21)	0.80
Umigon	4.71 (±1.48)	0.79
Semantria	4.78 (±2.12)	0.81
Panas-t	6.14 (±2.34)	0.79
AFINN	6.14 (±0.72)	0.78
Google SA	7.07 (±1.81)	0.76
IBM Watson	7.25 (±9.18)	0.73
OpinionLexicon	8.07 (±1.06)	0.73
MPQA	9.00 (±1.17)	0.73
Emolex	10.21 (±0.83)	0.70
Stanford	11.14 (±2.07)	0.66
ML-Sentistrength	11.40 (±1.45)	0.69
NRCHashtag	13.00 (±1.00)	0.62
SASA	13.50 (±1.00)	0.61
Happiness Index	14.21 (±0.53)	0.58

Table 4.4. Average Ranking using Macro-F1 as positional metric

top. If you consider both metrics Google Sentiment Analysis API has a great advantage, it has a Macro-F1 only 0.07 behind the best method (Emoticons) and has almost a perfect Applicability . Second, ten of our 15 shows better results than the best native method (ML-Sentistrength) for Applicability .

In summary, our results show that native methods do not administer well the trade-off between Macro-F1 and Applicability . This assumption can be verified at Table A.1 to Table A.14 , wherein many datasets, for example, French (Semantria), Portuguese (Semantria), English (Semantria, IBM Watson), Greek (ML-Sentistrength), these methods have a Applicability below 0.6. Also, we show that SOCAL and Sentistrength are better for both metrics compared to all native methods. This ultimate result provides evidence that our hypothesis is valid, English state-of-the-art sentiment analysis methods combined machine translator systems can be as good, or even better than "off-the-shelf" native methods. This result triggers an alert to authors from these methods which should compare their methods not only with other native methods but also the baseline proposed in this work.

Table 4.5. Average ranking considering **Applicability**

Method Name	**Average Ranking**	**Mean Applicability**
Google SA	0.71 (± 0.29)	0.98
NRCHashtag	0.79 (± 0.42)	0.98
Stanford	2.43 (± 0.40)	0.91
AFINN	4.86 (± 0.49)	0.76
Sentistrength	5.21 (± 1.16)	0.77
Emolex	5.50 (± 1.21)	0.75
SASA	5.64 (± 1.92)	0.80
SOCAL	6.79 (± 0.53)	0.73
OpinionLexicon	7.64 (± 0.58)	0.70
Happiness Index	8.71 (± 0.99)	0.67
ML-Sentistrength	9.00 (± 3.63)	0.63
Umigon	9.07 (± 1.21)	0.65
IBM Watson	9.50 (± 6.41)	0.60
Vader	11.71 (± 0.65)	0.56
Semantria	12.11 (± 1.26)	0.50
MPQA	12.64 (± 0.54)	0.50
Panas-t	14.79 (± 0.59)	0.06
Emoticons	14.82 (± 0.68)	0.11

Table 4.6. Average Winning Points using fcov as positional metric

Chapter 5

iFeel System

We presented a technique for multilingual sentiment analysis and compared it with approach against native solutions. We also described how methods developed for English text, with the help of machine translators, can be as good as methods engineered specifically for a certain language. Thus, we want to make no just the methodology, but the whole set of methods easily available for other in the scientific community.

In this context, we propose iFeel 3.0,[1] a benchmark system for sentence-level multilingual sentiment analysis. First published at [Araújo et al., 2014], iFeel implemented only eight methods without multilingual support. On its second version published at [Araújo et al., 2016] we increased the set of methods to 19 and also introduced the multilingual approach presented in this work. Despite both previous publications of the system, due the high acceptance and use from the scientific community we decided to rebuild iFeel, now on its third version.

The main reason for the development of iFeel's third version was the scalability and stability not provided by the both previous versions. The system had a high peak of 100 users created, and due its high computational resources demands, when few users upload files to be analyzed in parallel it used to crash. Additionally, iFeel 2.0 was develop using the Meteor Framework [2], a NodeJs based framework for fast development and prototyping. However, due Meteor constantly changes, updates and deprecated libraries the manage of the previous iFeel 2.0 tool was unsustainable. So, we choose to recreate iFeel on Spring Framework[3] environment, using Java as main programming language. We chose Spring because its stability, and it was meant to support "in production" applications. Also, as Java being a statically typed language, it gives us many advantages such as earlier detec-

[1] iFeel is hosted on http://www.ifeel.dcc.ufmg.br

[2] https://www.meteor.com/

[3] https://spring.io/

tion of programming mistakes and a more robust IDE, compared to previous Python and Javascript iFeel versions.

5.1 iFeel Architecture and Functionalities

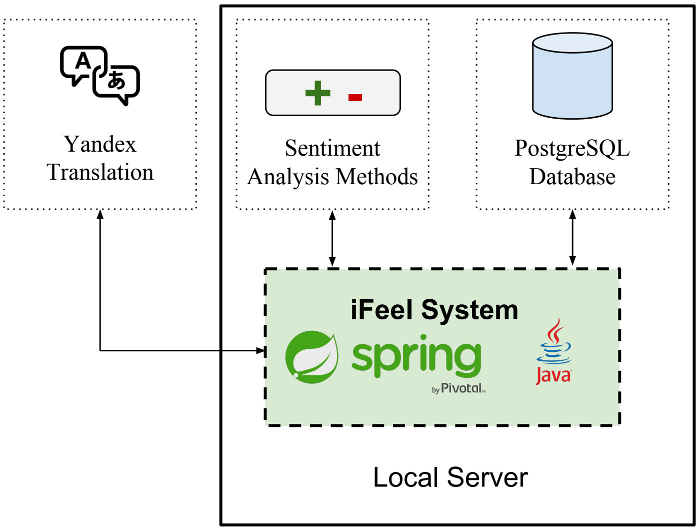
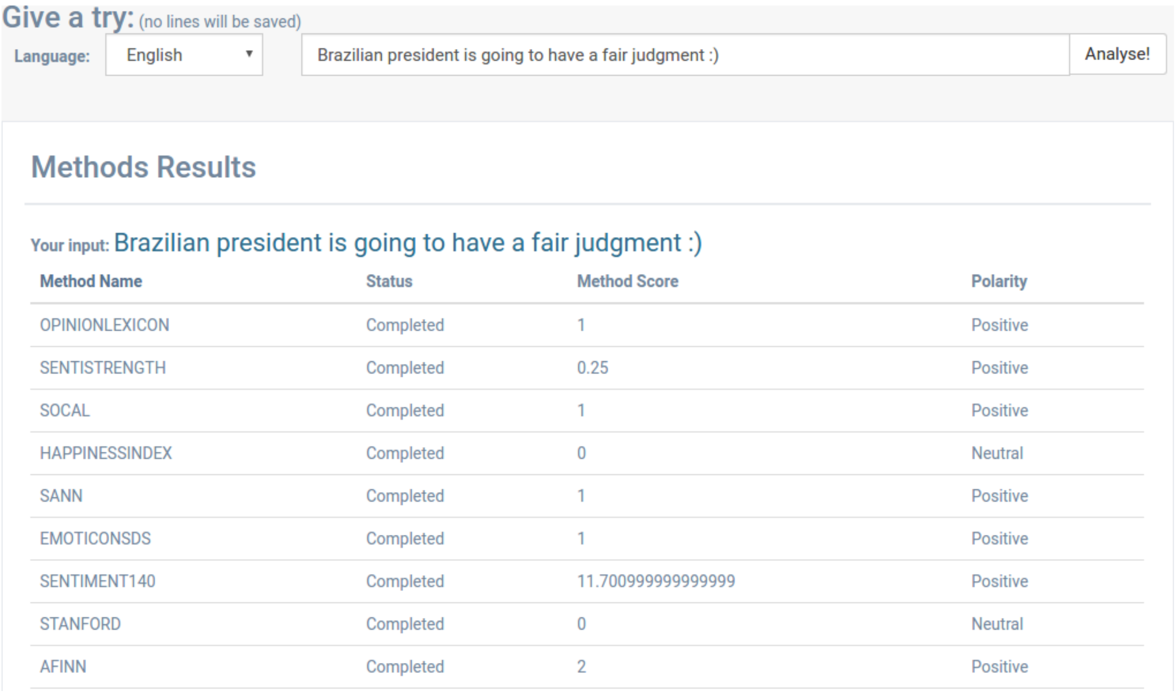


Figure 5.1. iFeel Architeture

The architecture of iFeel is represented in Figure 5.1. The local server runs the iFeel System implemented on the Spring Framework; it is responsible for the security layer, and view layer where the user can interact with the system. When the iFeel needs to perform sentiment analysis on sentences, it runs the Java version of the implemented methods available which can be download freely at `https://bitbucket.org/matheusaraujo/methodsjava`. IFeel is also connected to PostSQL database responsible for saving the sentences uploaded and also data from registered users. Finally, to perform multilingual sentiment analysis iFeel uses the Yandex Translate API and the approach defended in this work. It was chosen because it has the largest free tier among the top commercial machine translator systems.

In the first page of iFeel's interface, we present an introductory text along with our goals and functionalities, also, we want the user to test the system on their first contact. So, we leave two fields to be filled by the user, the language option, and a free text field when the form is submitted iFeel will perform the sentiment analysis polarity in all methods implemented as shown in figure 5.2. In the example, we submitted the text "Brazilian president is going to have a fair judgment :)" with the "English" language selected. We can see that most of the methods pointed the sentence as "positive", only the method



Brazilian president is going to have a fair judgment :)

Methods Results

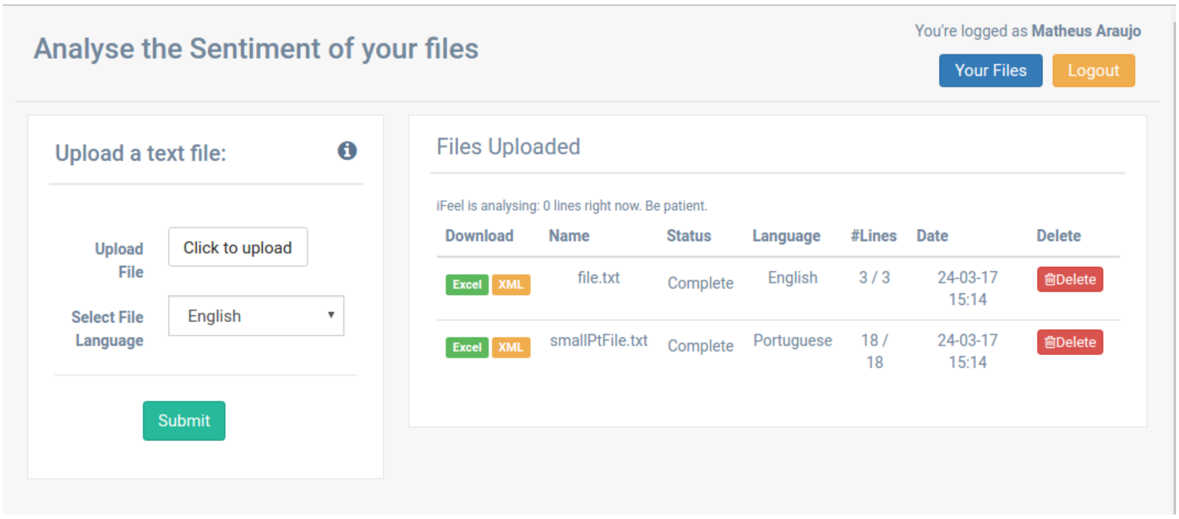
Your input: Brazilian president is going to have a fair judgment :)

Method Name	Status	Method Score	Polarity
OPINIONLEXICON	Completed	1	Positive
SENTISTRENGTH	Completed	0.25	Positive
SOCAL	Completed	1	Positive
HAPPINESSINDEX	Completed	0	Neutral
SANN	Completed	1	Positive
EMOTICONSDS	Completed	1	Positive
SENTIMENT140	Completed	11.700999999999999	Positive
STANFORD	Completed	0	Neutral
AFINN	Completed	2	Positive

Figure 5.2. iFeel - First user experience

Stanford and Happiness Index classified as "neutral". After the users register themselves, they have access to the "Analyse File Texts" page, where the user can upload multiple sentences in a file and trigger the iFeel system to analyze lines one by one. The upload page is shown in Figure 5.3. First, the user has to choose the language option (English by default). Then he has to upload the sentences from a plain text file, iFeel will perform a sentiment analysis for each line of the file with a maximum of 5000 sentences. The result is a *.xml* or *.xlsx* file which the user can download containing the output of all methods implemented.

A future step for iFeel is to provide a REST API for it's users. The ability of use iFeel automatically though an API was by far the must requested functionality by our users. It meets the need of the current state of Internet where microservices implemented for a machine-to-machine communication provided specialized functionality to be part of some larger solution. iFeel will always be free for scientific use, but it is also planned a commercial version of iFeel with paid features and technical support.



Analyse the Sentiment of your files

You're logged as **Matheus Araujo**

Upload a text file:

Upload File: Click to upload

Select File Language: English

Files Uploaded

iFeel is analysing: 0 lines right now. Be patient.

Download	Name	Status	Language	#Lines	Date	Delete
Excel XML	file.txt	Complete	English	3 / 3	24-03-17 15:14	Delete
Excel XML	smallPtFile.txt	Complete	Portuguese	18 / 18	24-03-17 15:14	Delete

Figure 5.3. iFeel - File upload section

Chapter 6

Conclusion

The Sentiment analysis field is currently popular and important to understand the social interactions throughout the Internet. People, companies, and even government agencies are using it to mine opinion inside digital forums, marketplaces, and social networks. The field has a certain value for Academy and commercial application. However, it is still limited by non-English content, not only in methods and tools but methodologies of how to solve the problem. Therefore, in this dissertation, we explored the issue of sentence-level multilingual sentiment analysis. Specifically, we analyzed how the current state-of-the-art English methods with the help of machine translators can solve this problem compared to previously published native methods.

First of all, we analyzed if choosing machine translators can affect the overall results of our experiments. To do so, we compared the results of translation-based methods using 3 different machine translators tools. Our conclusion regarding this topic is that machine translators are stable, showing consistent results among them all. Then, we present the results for English and native methods throughout all datasets, analyzing their performance related to Applicability , Macro-F1 and F1-score. We find that both approaches can detect positive sentences slightly better than negative sentences. We grouped the English methods and native methods and verify which approach is better, comparing the average Macro-F1 and Applicability across datasets. Then, using the average position across the languages datasets to ranking these methods, our findings suggest that the automatic translation of the input from a non-English language to English and the subsequent analyze in English methods can be a competitive strategy if the suitable sentiment analysis method is properly chosen. Moreover, we would recommend use the SOCAL or Sentistrength methods with the machine translation approach when analysing multilanguage texts. Moreover, we recommend using the SOCAL or Sentistrength methods with the machine translation approach when analyzing multilanguage texts.

Throughout this work, we presented many tentatives to implement multilingual sentiment analysis from the literature. However, our approach distinguishes from others in several ways. It is the first to analyze such wide variety of different languages with gold standard datasets. Additionally, the results show that machine translation aproach is a generic methodology that can be used in all languages supported by any proper machine translator.

We believe in two main direct application of this work. First, given the simplicity that the strategy of machine translation offers, we give a scientific foundation for who may prefer to deploy a multilingual sentiment analysis application at a small cost on instead of developing a solution on each particular language. Second, we hope that machine translation methodology could become a baseline for comparison of any novel language specific method.

As a final contribution, we provide the iFeel 3.0 system. Now, a more stable and reliable sentiment analysis framework developed using Spring. It implements many of the methods used in this work including a multilingual analysis support. We also release to the scientific community all the methods codes and labeled datasets used in this paper hoping that it can help sentiment analysis to become English independent.

Bibliography

Abdel-Hady, M., Mansour, R., and Ashour, A. (2014). Cross-lingual twitter polarity detection via projection across word-aligned corpora. *ICML WISDOM 2014 Conference*.

Abdulla, N., Ahmed, N., Shehab, M., and Al-Ayyoub, M. (2013). Arabic sentiment analysis: Lexicon-based and corpus-based. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on*, pages 1–6.

Al-Ayyoub, M., Essa, S. B., and Alsmadi, I. (2015). Lexicon-based sentiment analysis of arabic tweets. *International Journal of Social Network Mining*, 2(2):101--114.

Araújo, M., Diniz, J. P., Bastos, L., Soares, E., Ferreira, M., Ribeiro, F., and Benevenuto, F. (2016). ifeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis. In *Tenth International AAAI Conference on Web and Social Media*.

Araújo, M., Gonçalves, P., and Benevenuto, F. (2013). Measuring sentiments in online social networks. In *Proceedings of the 19th Brazilian symposium on Multimedia and the web*, pages 97--104. ACM.

Araújo, M., Gonçalves, P., Cha, M., and Benevenuto, F. (2014). ifeel: a system that compares and combines sentiment analysis methods. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 75--78. ACM.

Araujo, M., Nascimento, I., Rafael, G. C., de Melo-Minardi, R., and Benevenuto, F. (2016). Emotional fingerprint from authors in classical literature. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 263--270. ACM.

Araújo, M., Reis, J. C., Pereira, A. C., and Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *ACM Symposium on Applied Computing*.

Arora, P. (2013). *Sentiment Analysis for Hindi Language*. PhD thesis, Citeseer.

Bader, B. W., Kegelmeyer, W. P., and Chew, P. A. (2011). Multilingual sentiment analysis using latent semantic indexing and machine learning. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 45–52.

Balahur, A. and Turchi, M. (2012). Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 52--60. Association for Computational Linguistics.

Balahur, A. and Turchi, M. (2013). Improving sentiment analysis in twitter using multilingual machine translated data. In *RANLP*, pages 49--55.

Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127--135. Association for Computational Linguistics.

Basile, P. and Novielli, N. (2014). Uniba at evalita 2014-sentipolc task: Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. *Proceedings of EVALITA*, pages 58--63.

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

Bollen, J., Pepe, A., and Mao, H. (2009). Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. *CoRR*, abs/0911.1583.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida.

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.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y. A., Gelbukh, A., and Zhou, Q. (2016). Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Cognitive Computation*, 8(4):757--771.

Demirtas, E. and Pechenizkiy, M. (2013). Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pages 9:1--9:8, New York, NY, USA. ACM.

Dodds, P. S. and Danforth, C. M. (2009). Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82--89.

Ghorbel, H. and Jacot, D. (2011). Further experiments in sentiment analysis of french movie reviews. In *Advances in Intelligent Web Mastering–3*, pages 19--28. Springer.

Glavaš, G., Korenčić, D., and Šnajder, J. (2013). Aspect-oriented opinion mining from user reviews in croatian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 18--23, Sofia, Bulgaria. Association for Computational Linguistics.

Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on online social networks*, pages 27--38. ACM.

Goncalves, P., Benevenuto, F., and Cha, M. (2013). PANAS-t: A Pychometric Scale for Measuring Sentiments on Twitter. abs/1308.1857v1.

Gonçalves, P., Dalip, D. H., Costa, H., Gonçalves, M. A., and Benevenuto, F. (2016). On the combination of off-the-shelf sentiment analysis methods. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1158--1165. ACM.

Google (2017). Creating a sentiment analysis model.

Hai, Z., Chang, K., and Kim, J.-j. (2011). Implicit feature identification via co-occurrence association rule mining. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'11, pages 393--404, Berlin, Heidelberg. Springer-Verlag.

Hannak, A., Anderson, E., Barrett, L. F., Lehmann, S., Mislove, A., and Riedewald, M. (2012). Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*.

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.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. KDD '04, pages 168--177.

Hutto, C. 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*.

IBM (2017). Sentiment analysis with alchemyapi: A hybrid approach.

Khan, A. Z., Atique, M., and Thakare, V. (2015). Combining lexicon-based and learning-based methods for twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, page 89.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kincl, T., Novák, M., and Přibil, J. (2013). Getting inside the minds of the customers: automated sentiment analysis. In *European Conference on Management Leadership and Governance ECMLG*, pages 122--129.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *J. Artif. Int. Res.*, 50(1):723--762.

Koltsova, O. Y., Alexeeva, S., and Kolcov, S. (2016). An opinion word lexicon and a training dataset for russian sentiment analysis of social media. pages 277--287. -.

Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(24):8788--90.

Levallois, C. (2013). Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 414--417, Atlanta, Georgia, USA. Association for Computational Linguistics.

Lewis, W. (2010). Haitian creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *EAMT 2010: Proceedings of the 14th Annual conference of the European Association for Machine Translation, Saint-Raphaël, France. 8pp.*

Lexalytics (2017). Semantria api.

Li, S., Ju, S., Zhou, G., and Li, X. (2012). Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 139--148. Association for Computational Linguistics.

Lin, Z., Tan, S., and Cheng, X. (2011). Language-independent sentiment classification using three common words. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1041--1046. ACM.

Lo, S. L., Cambria, E., Chiong, R., and Cornforth, D. (2016). Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, pages 1--29.

Lu, B., Tan, C., Cardie, C., and Tsou, B. K. (2011). Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 320--330, Stroudsburg, PA, USA. Association for Computational Linguistics.

Makrynioti, N. and Vassalos, V. (2015). *Sentiment Extraction from Tweets: Multilingual Challenges*, pages 136--148. Springer International Publishing, Cham.

McKelvey, C. (2016). How the internet is changing the english language.

Meng, X., Wei, F., Liu, X., Zhou, M., Xu, G., and Wang, H. (2012). Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 572--581. Association for Computational Linguistics.

Messias, J., Diniz, J. P., Soares, E., Ferreira, M., Araujo, M., Bastos, L., Miranda, M., and Benevenuto, F. (2016). Towards sentiment analysis for mobile devices. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 1390--1391. IEEE.

Microsoft (2017). Machine translation.

min Kim, S. and Hovy, E. (2007). Crystal: Analyzing predictive opinions on the web. In *In EMNLPCoNLL 2007*.

Mohammad, S. (2012). #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246--255, Montréal, Canada. Association for Computational Linguistics.

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

Narr, S., Hulfenhaus, M., and Albayrak, S. (2012). Language-independent twitter sentiment analysis. *Knowledge discovery and machine learning (KDML), LWA*, pages 12--14.

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. (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 *ACL Conference on Empirical Methods in Natural Language Processing*, pages 79--86.

Pannala, N. U., Nawarathna, C. P., Jayakody, J. T. K., Rupasinghe, L., and Krishnadeva, K. (2016). Supervised learning based approach to aspect based sentiment analysis. In *2016 IEEE International Conference on Computer and Information Technology (CIT)*, pages 662–666.

Plutchik, R. (1980). *A general psychoevolutionary theory of emotion*, pages 3--33. Academic press, New York.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339--346, Stroudsburg, PA, USA. Association for Computational Linguistics.

Refaee, E. and Rieser, V. (2015). Benchmarking machine translated sentiment analysis for arabic tweets. In *HLT-NAACL*, pages 71--78.

Reis, J., Benevenuto, F., Vaz de Melo, P., Prates, R., Kwak, H., and An, J. (2015a). Breaking the news: First impressions matter on online news. In *Proceedings of the 9th International AAAI Conference on Web-Blogs and Social Media*, Oxford, UK.

Reis, J., Goncalves, P., Vaz de Melo, P., Prates, R., and Benevenuto, F. (2014). Magnet news: You choose the polarity of what you read. In *International AAAI Conference on Web-Blogs and Social Media*.

Reis, J. C., Gonçalves, P., Araújo, M., Pereira, A. C., and Benevenuto, F. (2015b). Uma abordagem multilıngue para análise de sentimentos. In *IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015)*.

Remus, R., Quasthoff, U., and Heyer, G. (2010). Sentiws-a publicly available german-language resource for sentiment analysis. In *LREC*.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., and Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1--29.

Rosenthal, S. and Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter.

Ruder, S., Ghaffari, P., and Breslin, J. G. (2016). INSIGHT-1 at semeval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis. *CoRR*, abs/1609.02748.

Ríos, A. A., Amarilla, P. J., and Lugo, G. A. G. (2014). Sentiment categorization on a creole language with lexicon-based and machine learning techniques. In *2014 Brazilian Conference on Intelligent Systems*, pages 37–43.

Shalunts, G., Backfried, G., and Commeignes, N. (2016). The impact of machine translation on sentiment analysis. *DATA ANALYTICS 2016*, page 63.

Shannon Greenwood, Andrew Perrin, M. D. (2016). Demographics of social media users in 2016. `http://www.pewinternet.org/2016/11/11/social-media-update-2016/`. Accessed in May, 28, 2017.

Shi, H. X. and Li, X. J. (2011). A sentiment analysis model for hotel reviews based on supervised learning. In *2011 International Conference on Machine Learning and Cybernetics*, volume 3, pages 950–954.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., and Gordon, J. (2013). Empirical study of machine learning based approach for opinion mining in tweets. In *Proceedings of the 11th Mexican International Conference on Advances in Artificial Intelligence - Volume Part I*, MICAI'12, pages 1--14, Berlin, Heidelberg. Springer-Verlag.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631--1642.

Souza, M. and Vieira, R. (2012). Sentiment analysis on twitter data for portuguese language. In *Computational Processing of the Portuguese Language*, pages 241--247. Springer.

Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 70--74, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Tamersoy, A., De Choudhury, M., and Chau, D. H. (2015). Characterizing smoking and drinking abstinence from social media. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media (HT).*

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *J. of Lang. and Soc. Psych.*, 29.

Thelwall, M. (2013). Heart and soul: Sentiment strength detection in the social web with sentistrength. `http://sentistrength.wlv.ac.uk/documentation/SentiStrengthChapter.pdf`.

Tromp, E. (2012). *Multilingual sentiment analysis on social media.* Lap Lambert Academic Publ.

Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, 24(3):478--514.

V. Le, Q. and Schuster, M. (2016). A neural network for machine translation, at production scale.

Vilares, D., Alonso, M. A., and Gmez-Rodrguez, C. (2017). Supervised sentiment analysis in multilingual environments. *Inf. Process. Manage.*, 53(3):595--607.

Villena Román, J., Lana Serrano, S., Martínez Cámara, E., and González Cristóbal, J. C. (2013). Tass-workshop on sentiment analysis at sepln.

Wan, X. (2008). Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 553--561, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *ACL System Demonstrations.*

Watson, D. and Clark, L. (1985). Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology*, 54(1):1063–1070.

Wikipedia (2017). Semantria — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Semantria&oldid=770334302`. [Online; accessed 15-May-2017].

Wu, Y., Zhang, Q., Huang, X., and Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1533--1541, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yandex (2017). Machine translation.

Yang, C.-S. and Shih, H.-P. (2012). A rule-based approach for effective sentiment analysis. In *PACIS*, page 181.

Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 129--136, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yussupova, N., Bogdanova, D., and Boyko, M. (2012). Applying of sentiment analysis for texts in russian based on machine learning approach. In *IMMM 2012, The Second International Conference on Advances in Information Mining and Management*, pages 8--14.

Appendix A

Appendices

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.84	0.93	0.91	0.92	SOCAL
0.76	0.91	0.86	0.89	Semantria
0.94	0.87	0.87	0.87	Stanford
0.64	0.88	0.79	0.84	Vader
0.99	0.86	0.83	0.84	Google SA
0.79	0.84	0.81	0.82	Sentistrength
0.69	0.8	0.79	0.79	MPQA
0.87	0.82	0.7	0.76	AFINN
0.89	0.81	0.71	0.76	Emolex
0.80	0.76	0.76	0.76	Umigon
0.13	0.77	0.64	0.71	Panas-t
0.85	0.79	0.57	0.68	OpinionLexicon
0.97	0.58	0.72	0.65	NRCHashtag
0.82	0.7	0.45	0.57	SASA
0.86	0.73	0.33	0.53	Happiness Index
0.00	0	0	0.00	Emoticons
-	-	-	-	IBM Watson
-	-	-	-	ML-Sentistrength

Table A.1. Simplified Chinese

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.15	0.98	0.91	0.94	Emoticons
0.74	0.88	0.80	0.84	Umigon
0.65	0.87	0.78	0.82	IBM Watson
0.39	0.88	0.73	0.80	Semantria
0.79	0.84	0.74	0.79	Sentistrength
0.74	0.83	0.74	0.78	SOCAL
0.60	0.87	0.69	0.78	Vader
0.98	0.85	0.70	0.77	Google SA
0.74	0.81	0.68	0.75	AFINN
0.06	0.82	0.67	0.74	Panas-t
0.74	0.79	0.64	0.72	Emolex
0.70	0.79	0.64	0.72	OpinionLexicon
0.50	0.75	0.68	0.72	MPQA
0.63	0.84	0.58	0.71	ML-Sentistrength
0.75	0.76	0.61	0.69	SASA
0.92	0.61	0.66	0.64	Stanford
0.98	0.63	0.64	0.64	NRCHashtag
0.68	0.75	0.39	0.57	Happiness Index

Table A.2. German

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.05	0.96	0.89	0.92	Panas-t
0.52	0.96	0.86	0.91	Vader
0.54	0.91	0.83	0.87	Semantria
0.77	0.91	0.83	0.87	Sentistrength
0.75	0.91	0.82	0.86	SOCAL
0.80	0.89	0.77	0.83	AFINN
0.03	0.98	0.67	0.82	Emoticons
0.72	0.88	0.72	0.80	OpinionLexicon
0.57	0.89	0.69	0.79	Umigon
0.49	0.85	0.74	0.79	MPQA
0.62	0.91	0.63	0.77	IBM Watson
0.78	0.84	0.67	0.76	Emolex
0.98	0.83	0.63	0.73	Google SA
0.64	0.84	0.47	0.66	Happiness Index
0.94	0.58	0.62	0.60	Stanford
0.99	0.55	0.59	0.57	NRCHashtag
0.66	0.78	0.35	0.57	SASA
-	-	-	-	ML-Sentistrength

Table A.3. Spanish

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.79	0.79	0.85	0.82	Sentistrength
0.52	0.83	0.81	0.82	Vader
0.61	0.79	0.83	0.81	Umigon
0.76	0.79	0.82	0.81	SOCAL
0.78	0.76	0.8	0.78	AFINN
0.75	0.76	0.78	0.77	OpinionLexicon
0.05	0.69	0.84	0.77	Panas-t
0.51	0.7	0.8	0.75	MPQA
0.04	0.91	0.51	0.71	Emoticons
0.93	0.6	0.81	0.71	Stanford
0.81	0.69	0.71	0.70	Emolex
0.98	0.7	0.71	0.70	Google SA
0.45	0.69	0.63	0.66	ML-Sentistrength
0.71	0.61	0.63	0.62	SASA
0.99	0.47	0.76	0.61	NRCHashtag
0.66	0.65	0.55	0.60	Happiness Index
-	-	-	-	Semantria
-	-	-	-	IBM Watson

Table A.4. Greek

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.05	0.96	0.98	0.97	Panas-t
0.59	0.89	0.83	0.86	Vader
0.79	0.85	0.81	0.83	Sentistrength
0.12	0.90	0.76	0.83	Emoticons
0.72	0.82	0.81	0.82	SOCAL
0.68	0.83	0.79	0.81	Umigon
0.75	0.83	0.78	0.80	AFINN
0.97	0.82	0.77	0.79	Google SA
0.54	0.79	0.75	0.77	Semantria
0.74	0.79	0.73	0.76	Emolex
0.72	0.79	0.72	0.75	OpinionLexicon
0.51	0.72	0.73	0.73	MPQA
0.74	0.68	0.68	0.68	ML-Sentistrength
0.98	0.62	0.72	0.67	NRCHashtag
0.71	0.73	0.56	0.65	SASA
0.66	0.75	0.55	0.65	Happiness Index
0.93	0.52	0.71	0.62	Stanford
0.27	0.00	0.87	0.43	IBM Watson

Table A.5. French

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.20	0.99	0.82	0.90	Emoticons
0.89	0.95	0.79	0.87	SOCAL
0.99	0.93	0.76	0.84	Google SA
0.91	0.94	0.72	0.83	Sentistrength
0.84	0.95	0.71	0.83	Vader
0.91	0.93	0.69	0.81	AFINN
0.91	0.91	0.55	0.73	Emolex
0.86	0.91	0.54	0.72	OpinionLexicon
0.89	0.85	0.59	0.72	Umigon
0.06	0.84	0.57	0.71	Panas-t
0.95	0.85	0.57	0.71	Stanford
0.72	0.83	0.56	0.70	MPQA
0.80	0.84	0.49	0.66	SASA
0.99	0.67	0.5	0.58	NRCHashtag
0.87	0.88	0.26	0.57	Happiness Index
-	-	-	-	Semantria
-	-	-	-	IBM Watson
-	-	-	-	ML-Sentistrength

Table A.6. Croatian

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.35	0.91	0.83	0.87	Vader
0.78	0.87	0.82	0.84	SOCAL
0.38	0.83	0.8	0.82	MPQA
0.38	0.79	0.76	0.78	Umigon
0.71	0.82	0.72	0.77	OpinionLexicon
0.58	0.78	0.75	0.77	Sentistrength
0.63	0.81	0.7	0.75	AFINN
1.00	0.76	0.59	0.67	Google SA
0.91	0.62	0.68	0.65	Stanford
0.79	0.74	0.54	0.64	Emolex
0.05	0.87	0.4	0.63	Panas-t
0.79	0.71	0.54	0.62	SASA
0.63	0.74	0.31	0.53	Happiness Index
0.99	0.45	0.56	0.50	NRCHashtag
0.00	0	0	0.00	Emoticons
-	-	-	-	Semantria
-	-	-	-	IBM Watson
-	-	-	-	ML-Sentistrength

Table A.7. Hindi

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.78	0.92	0.88	0.90	Sentistrength
0.65	0.92	0.85	0.89	Semantria
0.68	0.93	0.83	0.88	Vader
0.82	0.89	0.83	0.86	AFINN
0.12	0.86	0.86	0.86	Panas-t
0.82	0.88	0.77	0.83	OpinionLexicon
0.77	0.84	0.78	0.81	Umigon
0.81	0.86	0.76	0.81	SOCAL
0.88	0.87	0.72	0.80	Emolex
0.61	0.83	0.77	0.80	MPQA
0.98	0.88	0.73	0.80	Google SA
0.98	0.75	0.7	0.73	NRCHashtag
0.84	0.81	0.53	0.67	Happiness Index
0.79	0.76	0.52	0.64	SASA
0.95	0.53	0.67	0.60	Stanford
0.00	0	0	0.00	Emoticons
-	-	-	-	IBM Watson
-	-	-	-	ML-Sentistrength

Table A.8. Dutch

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.78	0.71	0.82	0.77	SOCAL
0.39	0.66	0.86	0.76	Stanford
0.99	0.65	0.75	0.70	Google SA
0.85	0.63	0.75	0.69	Sentistrength
0.75	0.68	0.69	0.68	Vader
0.84	0.56	0.78	0.67	Umigon
0.90	0.63	0.68	0.65	AFINN
0.81	0.52	0.78	0.65	MPQA
0.83	0.53	0.72	0.62	SASA
0.91	0.58	0.6	0.59	Emolex
0.09	0.52	0.64	0.58	Panas-t
0.99	0.31	0.8	0.56	NRCHashtag
0.89	0.58	0.45	0.52	OpinionLexicon
0.88	0.53	0.35	0.44	Happiness Index
0.05	0.63	0.23	0.43	Emoticons
-	-	-	-	Semantria
-	-	-	-	IBM Watson
-	-	-	-	ML-Sentistrength

Table A.9. Czech

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.04	0.92	0.9	0.91	Emoticons
0.99	0.58	0.87	0.73	Google SA
0.29	0.63	0.78	0.71	Vader
0.58	0.5	0.76	0.63	AFINN
0.23	0.65	0.62	0.63	Umigon
0.35	0.54	0.64	0.59	OpinionLexicon
0.93	0.35	0.75	0.55	NRCHashtag
1.00	0.12	0.93	0.53	SASA
0.95	0.12	0.92	0.52	Stanford
0.00	0	1	0.50	Panas-t
0.42	0.44	0.56	0.50	SOCAL
0.27	0.4	0.54	0.47	MPQA
0.62	0.4	0.54	0.47	Sentistrength
0.37	0.3	0.58	0.44	Emolex
0.32	0.28	0.44	0.36	Happiness Index
-	-	-	-	Semantria
-	-	-	-	IBM Watson
-	-	-	-	ML-Sentistrength

Table A.10. Haitian Creole

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.88	0.92	0.88	0.90	IBM Watson
0.08	0.96	0.85	0.90	Emoticons
0.52	0.94	0.84	0.89	Vader
0.06	0.91	0.81	0.86	Panas-t
0.75	0.89	0.8	0.85	Sentistrength
0.62	0.89	0.82	0.85	SOCAL
0.63	0.88	0.8	0.84	Semantria
0.72	0.86	0.79	0.83	Umigon
0.72	0.87	0.77	0.82	AFINN
0.92	0.86	0.75	0.81	Google SA
0.66	0.84	0.69	0.77	OpinionLexicon
0.34	0.8	0.74	0.77	MPQA
0.61	0.81	0.69	0.75	Emolex
0.60	0.78	0.66	0.72	SASA
0.95	0.66	0.67	0.67	NRCHashtag
0.60	0.81	0.48	0.64	Happiness Index
0.82	0.55	0.66	0.60	Stanford
-	-	-	-	ML-Sentistrength

Table A.11. English

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.08	0.9	0.89	0.89	Emoticons
0.49	0.91	0.8	0.85	Vader
0.65	0.83	0.76	0.80	SOCAL
0.59	0.83	0.74	0.79	Semantria
0.68	0.83	0.72	0.78	AFINN
0.75	0.84	0.73	0.78	Sentistrength
0.56	0.82	0.74	0.78	Umigon
0.97	0.82	0.7	0.76	Google SA
0.60	0.81	0.7	0.75	OpinionLexicon
0.72	0.8	0.66	0.73	ML-Sentistrength
0.40	0.76	0.7	0.73	MPQA
0.67	0.79	0.65	0.72	Emolex
0.04	0.78	0.59	0.68	Panas-t
0.97	0.6	0.64	0.62	NRCHashtag
0.65	0.79	0.44	0.62	Happiness Index
1.00	0.51	0.6	0.56	SASA
0.92	0.46	0.63	0.55	Stanford
-	-	-	-	IBM Watson

Table A.12. Portuguese

Applicability	F1(+)	F1(-)	Macro-F1	Method Name
0.03	1	1	1.00	Emoticons
0.56	0.86	0.87	0.86	Vader
0.07	0.83	0.87	0.85	Panas-t
0.70	0.83	0.87	0.85	Umigon
0.81	0.83	0.85	0.84	Sentistrength
0.77	0.78	0.83	0.81	SOCAL
0.08	0.67	0.95	0.81	Semantria
0.82	0.77	0.83	0.80	AFINN
0.71	0.72	0.77	0.75	OpinionLexicon
0.52	0.7	0.78	0.74	MPQA
0.98	0.74	0.74	0.74	Google SA
0.76	0.71	0.72	0.71	Emolex
0.91	0.52	0.76	0.64	Stanford
0.99	0.48	0.75	0.62	NRCHashtag
0.70	0.67	0.53	0.60	Happiness Index
1.00	0.44	0.74	0.59	SASA
-	-	-	-	IBM Watson
-	-	-	-	ML-Sentistrength

Table A.13. Russian

Applicability	**F1(+)**	**F1(-)**	**Macro-F1**	**Method Name**
0.57	0.77	0.8	0.79	Umigon
0.04	0.93	0.63	0.78	Emoticons
0.04	0.76	0.76	0.76	Panas-t
0.77	0.72	0.79	0.76	Sentistrength
0.49	0.78	0.74	0.76	Vader
0.73	0.71	0.78	0.75	SOCAL
0.79	0.71	0.75	0.73	AFINN
0.51	0.69	0.76	0.73	MPQA
0.71	0.68	0.7	0.69	OpinionLexicon
0.89	0.54	0.81	0.68	Stanford
0.80	0.6	0.74	0.67	Emolex
0.35	0.65	0.66	0.66	Semantria
0.62	0.64	0.68	0.66	ML-Sentistrength
0.95	0.59	0.7	0.65	Google SA
0.98	0.46	0.78	0.62	NRCHashtag
0.63	0.61	0.5	0.55	Happiness Index
0.79	0.55	0.55	0.55	SASA
-	-	-	-	IBM Watson

Table A.14. Italian