Information Sciences 512 (2020) 1078–1102

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

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

Matheus Araújo[a,b,*], Adriano Pereira[b], Fabrício Benevenuto[b]

[a] University of Minnesota, Minneapolis, USA
[b] Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil

ARTICLE INFO

Article history:
Received 3 January 2018
Revised 8 October 2019
Accepted 14 October 2019
Available online 14 October 2019

Keywords:
Sentiment analysis
Multilingual
Machine translation
Online social networks
Opinion mining

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 identification of 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 other 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, and compared them with three language-specific methods. Based on fourteen human labeled language-specific datasets, we provide an extensive quantitative analysis of existing multilingual approaches. Our results suggest that simply translating the input text in a specific language to English and then using one of the existing best methods developed for English can be better than the existing language-specific approach evaluated. We also rank methods according to their prediction performance and identify those 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 and tool for multilingual sentence-level sentiment analysis[1]. We hope our system sets up a new baseline for future sentence-level methods developed in a wide set of languages.

1. Introduction

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 when

* Corresponding author.
E-mail addresses: arauj021@umn.edu (M. Araújo), adrianoc@dcc.ufmg.br (A. Pereira), fabricio@dcc.ufmg.br (F. Benevenuto).

[1] iFeel resources: https://sites.google.com/view/ifeel-resources/home.

https://doi.org/10.1016/j.ins.2019.10.031
0020-0255/

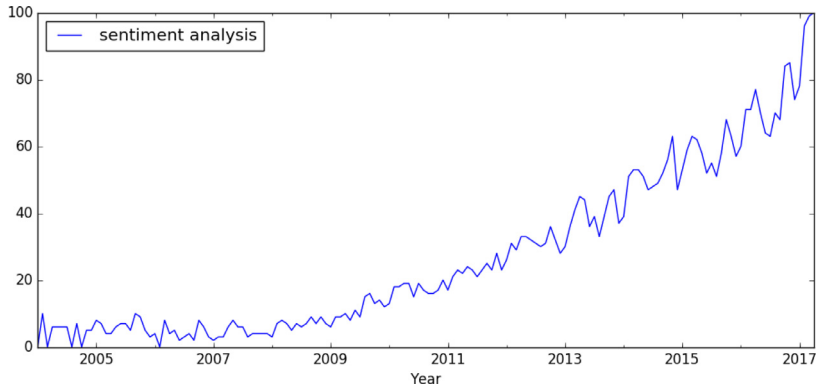

Fig. 1. Interest in "Sentiment Analysis" since 2004 according to Google Trends.

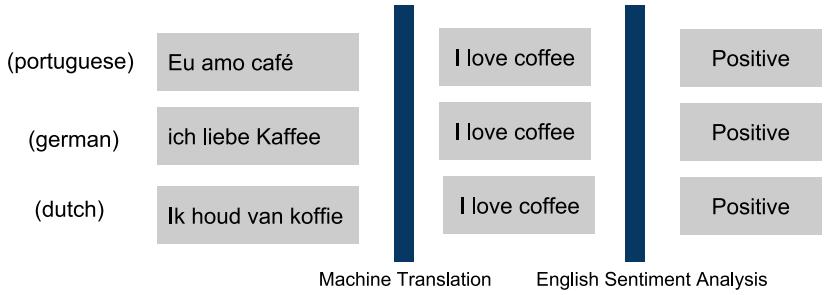

Fig. 2. Our methodology overview with a simple example.

investing in tools for massive marketing campaigns. Thus, sentiment analysis has become a hot topic in Web applications, with high demand from industry and academy. Fig. 1 demonstrates the rising popularity of sentiment analysis.

Despite the large interest from industry and academy in the sentiment analysis field, substantial effort has been focused only on the English idiom, since it is dominant across the Web content [32]. However, the potential market for sentiment analysis in different languages is vast. Until now, a mobile application that uses sentiment analysis in several countries, would require dealing with sentiment analysis approaches on multiple languages, which is currently quite limited. Some efforts even attempt to develop techniques to analyze sentiments from other specific languages: Arabic [30], German [27], Portuguese [38], Russian [16], Spanish [35], among others. However, little is known about the performance prediction, viability and real need of those methods. More important, a different solution for 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 translation strategy can address the problem of sentiment analysis in multiple languages. Additionally, it argues for the use of translation-based techniques as a baseline for new multilingual sentiment analysis methods. Particularly, it analyzes how the use of machine translation systems - such as Google Translate[2], Microsoft Translator Text API[3] (used by Bing Translator[4]) and Yandex Translate[5] - combined with English sentiment analysis methods can be comparable to methods created specifically to non-English texts. In addition, we deeply investigate the accuracy of 3 well-known machine translation system, including an offline translation method for a comparison baseline.

Using the output from machine translation tools, we evaluate the prediction performance of 13 English sentiment analysis methods across 14 different languages: Chinese, German, Spanish, Greek, Croatian, Hindi, Czech, Dutch, French, Haitian Creole, English, Portuguese, Russian, Italian. Then, we compare our analysis with the results derived from methods originally created to classify these languages. According to *Internet World Stats*[6], seven of these languages appear among the top ten languages used on the Web and represent more than 61% of non-English speaker users. Fig. 2 presents an overview of our methodology.

There is still a large space for improvement in the current state-of-the-art of sentiment analysis in English. This potential is suggested by a benchmark study [32], as well as recent improvements using deep neural networks, more specifically LSTMs [50]. However, our findings suggest that machine translation systems are mature enough to produce reliable trans-

[2] https://translate.google.com.
[3] https://www.microsoft.com/en-us/translator/translatorapi.aspx.
[4] https://www.bing.com/translator.
[5] https://translate.yandex.com/.
[6] http://www.internetworldstats.com/stats7.htm.

lations to English that can be used for sentence-level sentiment analysis and obtain a competitive prediction performance when compared to methods built for a specific language. Additionally, we show that some popular language-specific methods do not have a significant advantage over the machine translation approach.

The remainder of this work is organized as follows. Next section provides an overview of sentiment analysis and efforts for multiple-language approaches. Then, Section 3 describes our methodology for evaluating sentiment analysis methods in multiple languages. Section 4 presents comparative results among methods and machine translation approaches for multiple languages. Finally, Section 6 details iFeel systems and the new features related to multiple languages we deployed in the system. The last section summarizes our contributions and suggests directions for future work.

2. Sentiment analysis overview

The recent popularity of the term sentiment analysis has led to its use to describe a wide variety of tasks by the community. Therefore, a broad concept of sentiment analysis exists, 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 the Conference on Empirical Methods in Natural Language Processing (EMNLP). The annual SemEval[7] workshop stands out as one that evaluates the current state-of-the-art techniques and proposes several new challenging tasks for the field. In 2017, the SemEval workshop had five tasks related to sentiment analysis. Each of these tasks has many subtasks, ranging from three-class polarity detection of tweets to veracity prediction of a rumor.

In this study, we focus on the use of "off-the-shelf" methods to perform multilingual sentiment analysis. For granularity filter, we focus on sentence-level methods that particularly target the 2-class polarity problem (positive vs. negative).

Although many of the methods we use generates 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. [32]. However, for the purpose of this work, we simplify our experiments, focusing only on the 2-class problem.

2.1. Literature review on 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. For a comprehensive overview of existing methods designed to English we refer the reader to references [32] and [22], in which the authors compared many "off-the-shelf" methods and attempted to combine them in a single one.

In general, many multilanguage approaches focus on adapting strategies that previously succeeded for English to other languages.

2.1.1. Machine translation-based methods

Refaee and Rieser [30] 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. [35], the potential of machine translation on sentiment analysis is also explored, using the combination of two 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.

In [6], the authors investigate 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, [5] investigates the problem of sentiment detection in three different languages: French, German, and Spanish. Their focus was evaluate how an automatic translation of text would work to obtain training and test data for these three languages and subsequently extract features that were employed to build machine learning models using Support Vector Machines.

Nevertheless, our work is the first to test this technique in a wide range covering 14 different languages and comparing the results of 13 "off-the-shelf" English sentiment analysis methods against 3 language-specific methods to support the "simple translation" hypothesis.

2.1.2. Lexicon and corpus-based methods

In 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 or sentiment polarity from the respective rule. Consider the following extraction rules example [48]: lithium-ion -> battery, mAh -> battery, rechargeable -> battery; they indicate that if "lithium-ion," "mAh," or rechargeable,å appears

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

in a review sentence, we have high confidence 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 [44], the authors propose an approach that uses an English dataset to improve the results for a Chinese sentiment analysis using rule-based approach. The authors apply a set of lexicons to build rules that consider: 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 a class. In the same direction, Mihalcea et al. [24] proposes a rule-based approach to classify Romanian text by building a new subjectivity lexicon from translating an existing English one.

2.1.3. Machine learning-based methods

Many of the proposed methods uses at least in part machine learning techniques for multilanguage sentiment analysis (not limited to this subsection). 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 [19]. 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 data collection.

2.1.4. Parallel corpus-based methods

A particular approach for multilingual 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. Using 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 have the same sentiment. Therefore, their goal is to identify and maximize the joint likelihood of the language-specific labeled to infer its sentiment labels [20].

In [23], 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 in which the machine translators have a limited vocabulary. Their results show an improvement of 12% in the accuracy using this approach when combining corpus in English and Chinese.

A novel methodology using a parallel corpus is also proposed by Bader et al. [4]. In this case, the authors use different datasets from Bible translations in many 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.å In order to prevent a high dependency of the model given the Bible context, a step in their methodology was to shuffle the sentences, a technique that helps break any topic/sentiment association. Their results for accuracy ranges from 72% to 75% to non-English evaluations.

2.2. Deep neural networks methods applied to multilingual sentiment analysis

Deep Neural Networks, or deep learning-based methods, recently shows a promising approach for sentiment analysis field. This claim covers the wide range of tasks and multiple levels of granularity, which transform deep-learning in a strong candidate to become the state-of-the-art technique as discussed in by Zhang et al. [49]. One example is the use of convolutional neural network (CNN) for both: aspect extraction and aspect-based sentiment analysis proposed by Ruder et al. [33] in the SemEval-2016 Task 5 challenge. 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.

Besides CNNs, LSTMs (Long short-term memory) is another variation of deep-learning that gained recent popularity in tasks where natural language processing is required. In [50], the authors use a combination of Bidirectional LSTMs, CNN and max-pooling layers to build a neural network able to beat the state-of-the-art performance in 4 of 6 tasks. Similarly, Shuang et al. [36] uses a Bidirectional LSTM to build a Sentiment Information Collector and a Sentiment Information Extractor (SIE). The authors claim this approach is generic after testing using 3 datasets in English and Chinese.

Overall, although sentiment analysis is rich in solutions, it is still centered on the English context. In our effort, we compare many strategies and efforts for multilingual sentiment analysis. We understand that a comparison of "off-the-shelf" methods, associated with strategies of machine translation and applied to a wide range of languages is still missing in the literature and it has significant value for this research community, providing a baseline for comparison for future approaches.

3. Methodology

Our methodology to evaluate sentiment analysis in multiple languages by machine translation involves several key elements. The first element is a large set of sentiment analysis methods, designed for English text and able to identify if a

sentence is positive or negative. To obtain this set of methods, 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. The second key element is 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. The third key element is a baseline for comparison. This baseline is a set of sentiment analysis methods designed natively to non-English sentences that matches the languages in our dataset, described in Section 3.3. Finally, we test our hypothesis using 3 commercial machine translation systems and one word-by-word offline translator that were applied 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 [13], to gauge the news polarity [31]. Hence, we restrict our focus on those efforts related to detecting the polarity (i.e., positivity or negativity) of a given text. This can be done with small adaptations on the output of some existing methods, a methodology previously described by Araújo et al. [2], Gonçalves et al. [11].

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 their citations and those papers that cited them until the end of 2016. 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 1 presents an overview of the methods used in this work, the reference paper in which they were published and the main technique type that they are based on (ML - machine learning - or L - lexicon). As summarized in Table 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. [32]. Following their methodology, we choose 14 methods from that study.

Finally, we added the Google Prediction API in our set of sentiment analysis methods to analyze English texts. This method is a commercial sentiment analysis tool created by Google with respected accuracy. We added it to verify if there is a large discrepancy in the results 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 Chapter 6.

3.2. Human labeled datasets

In this section, we describe the multilingual datasets used to compare the sentiment analysis performance of machine-translation against native methods. These workloads consist of 14 gold standard datasets of sentences. Each sentence was labeled by humans as positive, negative or neutral according to their sentiment polarity. By using human annotations, we can compare the quality of the sentiment analysis methods and judge their performance. In Table 3 we summarize the relevant information about these datasets, showing in each row the language, its ISO 639-1 two-letter code, the publication it first appeared, which subtype of dataset if was collected, and the number of positive (Pos) and negative sentences(Neg)[8].

The process of acquiring these datasets was a crucial step in this work. It is very challenging to produce them because of two main reasons: the intrinsic subjectivity of each sentence (even cultural dependency) and a large amount of time needed to annotate them. In our case, we would have an extra challenge since humans who work in the annotation process should know fluently different languages. So, to successfully proceed with this work, we contacted various authors in the field who already did this labeling work in a specific language. The research policy applied considered all published paper that makes available its dataset and the sentences are labeled as positive, negative or neutral, either as disjoint classes or intensities of these 3 classes. The result of this extensive manual work is a unique and rich source of human-labeled sentences in many languages. After getting these 14 independent datasets, we post-process them to make sure that the labels are all in the same range. By doing so, we can compare human classification with the sentiment analysis methods output.

Note that not all of the human annotation tasks were made by the same research policies and standards. Some were labeled by three people whereas others were labeled by only two. Some used Amazon Mechanical Turkers and others used experts/specialists. Additionally, some datasets came from ramdom tweets in a language, but others used data focused on a specific theme. For instance, the Russian dataset was collected from product reviews in Russian blogs and the Croatian dataset that was came from food reviews. We perceive these differences in the decisions 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, the goal of lexical approaches for sentence-level sentiment analysis is usually to be more generic and independent of context as possible. Thus, we treat all the datasets equally without

[8] The datasets used in this paper are available under request at iFeel-resources (code and datasets) is available at https://homepages.dcc.ufmg.br/~fabricio/ifeel_resources.htm.

Table 1
List of sentiment analysis methods for English sentences obtained in the literature with their description and core techniques (L- Lexicon-based ML-Machine Learning-based).

Name	Description	L	ML
Emoticons([11])	Messages containing positive/negative emoticons are positive/negative. Messages without emoticons are not classified.	✓	
Opinion Lexicon [13]	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.	✓	
Opinion Finder (MPQA) [47]	Performs subjectivity analysis trough a framework with lexical analysis former and a machine learning approach latter.	✓	✓
Happiness Index [9]	Quantifies happiness levels for large-scale texts as lyrics and blogs. It uses ANEW words [8] to rank the documents.	✓	
AFINN [28] - A new ANEW	Builds a Twitter based sentiment Lexicon including Internet slangs and obscene words. AFINN can be considered as an expansion of ANEW [8], a dictionary created to provides emotional ratings for English words. ANEW dictionary rates words in terms of pleasure, arousal and dominance.	✓	
SO-CAL [39]	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 ([25]	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 [45]	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 [12]	Detects mood fluctuations of users on Twitter. The method consists of an adapted version (PANAS) Positive Affect Negative Affect Scale [46], 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 [26]	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 [29]. Proposed Lexicon includes unigrams and bigrams from Macquarie Thesaurus and also words from GI and Wordnet.	✓	
SentiStrength [40]	Builds a lexicon dictionary annotated by humans and improved with the use of Machine Learning.	✓	✓
Stanford Recursive Deep Model [37]	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 [17]	Disambiguates tweets using lexicon with heuristics to detect negations plus elongated words and hashtags evaluation.	✓	
VADER [14]	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 [a]	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.		✓

[a] Google Cloud Platform https://cloud.google.com/prediction/docs/sentiment_analysis, Date of Access: 2017-05-29.

Table 2
Sentiment analysis methods for English sentences and how their original output was mapped to positive (blue), neutral (black) and negative (red) classes.

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

Table 3
Summary of multilingual gold standard human labeled datasets.

Language	Neg	Pos	Published at	Code	subtype
Chinese	432	446	[44]	zh	product reviews
German	239	353	[27]	de	tweets
Spanish	350	683	[43]	es	tweets
Greek	3189	2131	[21]	el	tweets
French	321	341	[27]	fr	tweets
English	998	1595	[27]	en	tweets
Croatian	467	1658	[10]	hr	food reviews
Hindi	230	340	[3]	hi	product review
Dutch	43	77	[41]	nl	tweets
Czech	2808	1422	[15]	cs	movie reviews
Haitian Creole	734	128	[34]	ht	tweets
Portuguese	414	626	[27]	pt	tweets
Russian	416	333	[16]	ru	tweets
Italian	1422	820	[7]	it	product reviews

Table 4
Description of the sentence-level native methods used for comparison.

Name	Description	Paid
Semantria [a]	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.	✓
IBM Watson API (Alchemy API) [b]	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 [40]	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	

[a] Semantria API: https://www.lexalytics.com/semantria Date of Access: 2017-05-29.
[b] Sentiment analysis with alchemyapi: A hybrid approach. https://www.ibm.com/account/reg/us-en/signup?formid=mrs-form-287 Date of Access: 2017-05-29.

configuring or training the methods for a specific situation. This approach works for our main goal which is to test the following hypothesis: Are the machine translation to English and further analysis of sentiment by English-aimed methods as good as native methods analyzing texts directly in their languages?

3.3. Language-specific sentiment analysis methods (native methods)

Ideally, we want to compare the use of machine translation approach to English and the output of methods described in Section 3.1 with a large number of methods designed specifically for the dataset language. We contacted authors of several native methods. While we succeeded in obtaining a large number of datasets, most of these methods are not available even under request, making reproducibility almost impossible when comparing the original paper of the dataset.

At last, we were able to assess 3 “off-the-shelf” native methods created or trained specifically for certain languages to use as a baseline. In Table 4 we list and describe these methods shortly and in Table 5 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[9], which provides sentiment classification for sentences in 21 languages. We used the trial version of Semantria’s Microsoft Excel Plug-in available on their web site.

The third native method is the IBM Watson API, a commercial sentiment analysis toolkit developed by IBM, which has a range of features such as polarity detection, post-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 5 that 4 languages (i.e., Croatian, Hindi, Czech and Haitian Creole) do not have any native sentiment analysis method to compare with. Although the comparison results in this work refer to the languages that have at least one

[9] https://semantria.com.

Table 5
Reference of language support by popular multilingual sentiment analysis methods.

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

[a] Simplified/Standardized Chinese.

native method that supports it, we still show the results for all languages that we have access to human-labeled datasets. After all, we can still compare the performance between English methods from the machine translation approach. These results become a baseline for future authors who aim at developing native methods.

3.4. Machine translation systems

Since the 1950s, machine translation or automated translation is a field of research[10]. Its main goal is to provide text translation by a computer without human interaction. There are three main approaches to solve the problem of generating automatic translation: Rules-based/phrase-based, statistical methods or neural networks. The rules-based uses lexicons combined with grammar definitions to translate sentences in a meaningful way. The statistical system tries to build a translation model by analyzing a large amount of training data for each pair of languages. The neural networks based systems build one large artificial neural network by using a huge amount of training data, this approach has recently become popular and shows better translation performance. When considering the use of machine translation to perform multilingual sentiment analysis, we want to answer two important questions:

1. Why we choose commercial machine translators tools instead of free published tools?
2. Why we believe that machine-translated texts to English combined with English sentiment analysis tools are better than native non-English sentiment analysis methods?

To answer the first question, we need to clarify that there are available many free open sources machine translation tools for multiple languages[11]. However, these tools are based on a pre-trained statistical system, indeed they are static and do not follow the dynamic evolution of each language, especially on the Web. In other words, in environments such as the Internet, new emoticons, slangs or even ways to express are frequently generated, requiring constant training models[12]. So, we choose well-known commercial tools which retrain periodically their models, as explained by [42]. Since we do not have either resources or knowledge to keep an updated trained model of high accuracy in our environment and it is not the purpose of this work, we decided to use the commercial API's.

In Fig. 3 we see a comparison performance between three translators candidates, a neural network, a phrases-based system, and proper humans. The chart illustrates how close the current state-of-the-art machine translation systems are to humans translators. Also, it shows that neural networks seem to overcome the phrase-based strategy. So, we answer the second question by combining these results with the following axiom: words will probably change between two paired sentences in different languages, however, a reasonable machine translation should not change their sentiment polarity.

We used 3 popular commercial translation tools to translate our non-English datasets to English, they are listed in Table 6. About these tools, the Yandex API allows the user to send 10,000 free requests per month. Google Translator has a free Web interface but no free API support, but you are granted with U$300,00 when creating an account on the Google Cloud Platform, in which every million of characters translated will cost of U$20. The Microsoft Translator Text API can be used inside the Microsoft Azure platform, 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.

[10] A neural network for machine translation, at production scale. Date of Access: 2019-05-29. https://research.googleblog.com/2016/09/a-neural-network-for-machine.html.

[11] http://fosmt.org/.

[12] How the Internet is changing the English language. Date of Access: 2017-05-29. https://www.dailydot.com/parsec/dialects-of-internet-communities/.

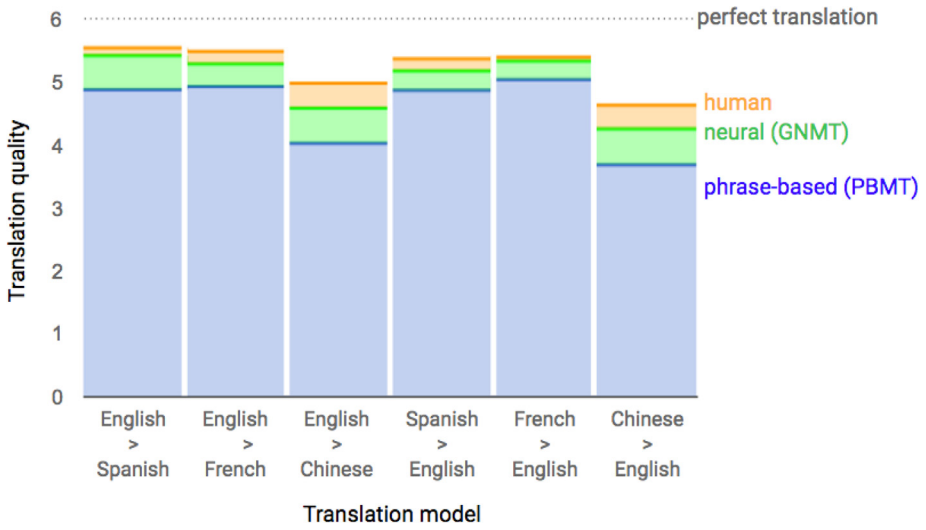

Fig. 3. Comparison between phrase-based and neural network techniques with a human baseline, extracted from [42].

Table 6
A summary of the machine translation tools used in this work to translate to English the desired text to be classified.

Translator	Description
Yandex Translate API [a]	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 [42]	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 [b]	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
Baseline	A word-by-word translator developed by us. Given each unique word in each language dataset, we translated the word individually without any contextual information.

[a] Yandex API: https://yandex.com/company/technologies/translation/ Date of Access: 2017-05-29.
[b] Microsoft Machine Translate: https://www.microsoft.com/en-us/translator/mt.aspx Date of Access: 2017-05-29.

Finally, we also compared the commercial machine translators results with a baseline method created by us. This baseline uses the most simple translation technique known, a word-by-word (literal translation). To build this translator, we first identified every each word in each language dataset. Using the Google Translator we translate each unique word, individually, to English. The final step was to substitute the words in the original dataset by its literal translation. This baseline was created to be a comparison step, where no contextual information is used in the translation process.

4. Experimental evaluation

In this section, we present all the experiments performed in this work to sustain the hypothesis that current sentiment analysis methods created for English combined with the current state-of-the-art machine translation system can classify sentiment in sentences as good as the native sentiment analysis methods.

Several experiments were performed to answer the following questions:

1. How choosing the machine translator impacts the overall performance?
2. What are the performance of English methods for sentiment analysis classification in non-English content when automatic machine translation is applied?
3. Is the machine-translated approach better than native methods?
4. Is there a difference of performance when positive and negative polarities are individually evaluated?
5. In which cases the native methods are better than the machine translation approach?

In the following subsection, we present the metrics we choose to compare the performance of the sentiment analysis techniques used in this work. After that, we show the experimental evaluation and a final summary of the results.

Table 7
Confusion matrix for positive and negative classes.

		Predicted	
		Positive	Negative
Actual	Positive	a	b
	Negative	c	d

4.1. Metrics

The **F1-Score** is a metric commonly used to compare the quality of the prediction given a ground truth. In our case, we use F1-Score to check how good a method can identify a sentiment in a sentence when compared to human annotation. The F1-Score considers equally important the precision and recall of the classification. This metric can be easily computed for 2-class experiments using the Table 7.

The precision of the positive class is computed as:

$$P(\text{positive}) = \frac{a}{(a+c)} \tag{1}$$

The recall is calculated as:

$$R(\text{positive}) = \frac{a}{(a+b)} \tag{2}$$

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

$$F1(\text{positive}) = \frac{2P(\text{positive}) \cdot R(\text{positive})}{P(\text{positive}) + R(\text{positive})} \tag{3}$$

A variation of the F1-Score named **Macro-F1** is normally reported to evaluate classification effectiveness on skewed datasets when the class distribution is not homogeneous. Hence, Macro-F1 is the metric we use in our analysis and it is computed by averaging the F1-Score for positive and negative classes. This metric considers equally important the effectiveness in *each class*, independently of the relative presence of the class in the dataset. In our analysis, we only considered a sentence to be used in the evaluation of a specific method if the method can indicate one of the 2-class, negative or positive. If a method output is neutral, then we ignore the sentence when computing the Macro-F1. Therefore, the Macro-F1 reported represents how effective the method is when it indicates that a sentiment polarity.

Although we only use the output of methods that indicates a sentiment polarity to compute Macro-F1, the methods still return the neutral class for several sentences. Thus, we define as **Applicability**, a metric that shows the percentage of sentences a method can 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 from the methods. Moreover, it seems that methods which are conservatives regarding given a polarity to sentence usually have 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 an important metric for a complete evaluation in the 2-class experiments. Even though Emoticons presents high accuracy, it was not able to predict 90% of the sentences. More formally, Applicability is calculated as:

$$\text{Applicability} = \frac{\text{total sentences} - \text{neutral sentences}}{\text{total sentences}} \tag{4}$$

In our analysis, we discuss the results and trade-off 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 when applied in multilingual sentiment analysis

In this section, we evaluate if there is a difference in the outcome results when choosing a specific machine translators system. We compared all 4 machine translators discussed previously to answer question 1. All the language datasets were translated from their original texts to English on these translators. An exception is the English dataset used only to be a comparison baseline.

In Figs. 4 and 5 we present the distribution performance overall datasets as boxplots when considering Macro-F1 and Applicability. For each machine translator, we compute results for both metrics overall English sentiment analysis methods output listed on Table 2. In respect to the Macro-F1, the distribution is very similar, especially between 25th and 75th percentile, with Google Translator slightly better than others. When averaging the Macro-F1 for all methods in all datasets, Yandex and Google machine translators have Macro-F1 of 0.73 with a standard deviation of 0.12. The Microsoft Translator has a marginally inferior performance among the commercial translators with a Macro-F1 average of 0.72 and a standard deviation

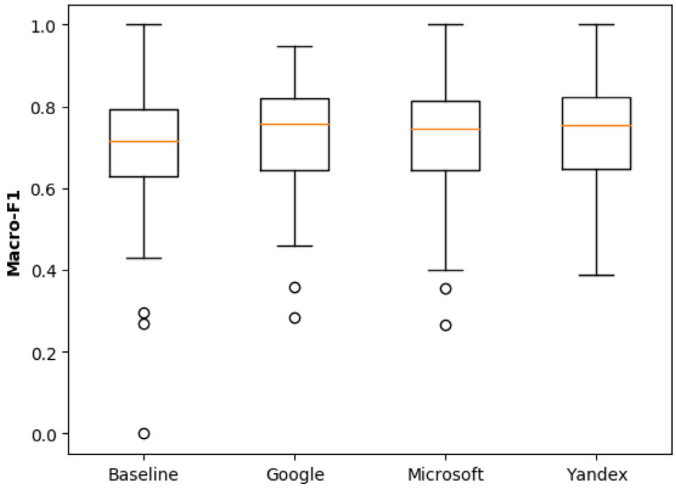

Fig. 4. Macro-F1 results given each machine translation system among all language datasets when translated to English.

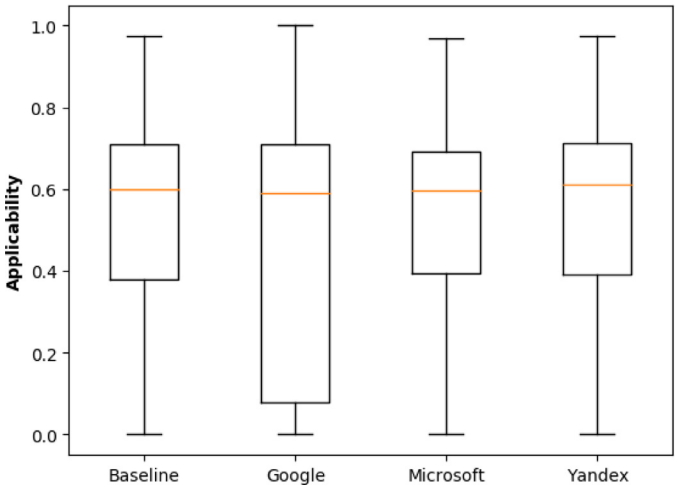

Fig. 5. Applicability results given each machine translation system among all language datasets when translated to English.

Table 8
Macro-F1 and Applicability mean ($\alpha = 0.95$) results given each machine translation system among all language datasets when translated to English.

Translation Method	Macro-F1	Applicability
Google	$0.73 \pm (0.02)$	$0.47 \pm (0.04)$
Yandex	$0.73 \pm (0.02)$	$0.54 \pm (0.03)$
Bing	$0.72 \pm (0.02)$	$0.53 \pm (0.03)$
Baseline	$0.70 \pm (0.02)$	$0.51 \pm (0.03)$

of 0.20. Despite this difference, the confidence intervals of all 3 commercial machine translators overlap for $\alpha = 0.95$, hence, no statistically significant differences were found. We had a similar result for Applicability, where choosing the translator do not statistically influence the result. In Table 8, we show an aggregate summary for each translator and metric. Although Yandex has the best performance when analyzing Macro-F1 and Applicability, it is still inside the confidence interval of others.

When comparing our baseline translator, the following findings should be highlighted. The baseline's Macro-F1 and Applicability results were inside the confidence intervals of the commercial translators, which is a great result for an offline translation method that does not consider contextual information of the sentence. There are many advantages to the baseline approach. First, it is easier and faster to translate sentences, since the method just looks at a hash dictionary the words to be translated. Second, since other translators charge the user per API access, having an offline method would be financially effective. The only disadvantage is the need for a large enough word dictionary that has to cover the input text words.

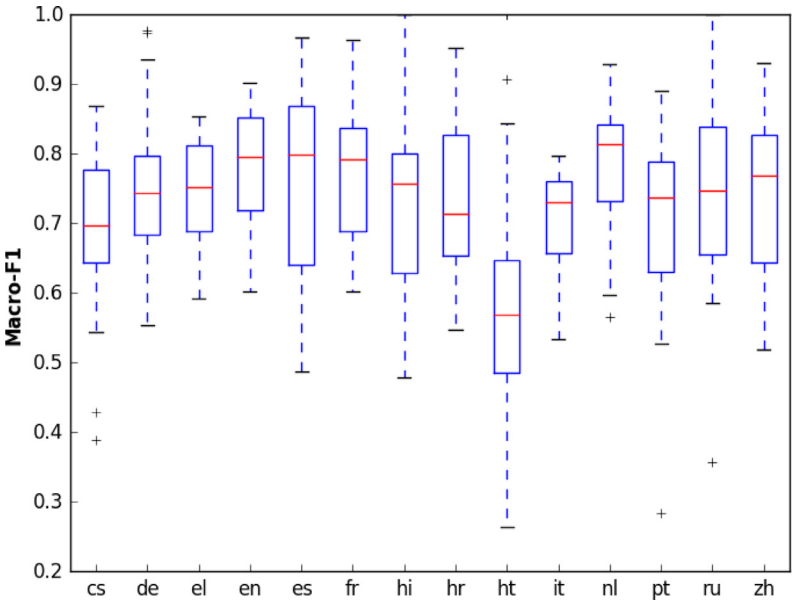

Fig. 6. Overall performance of sentiment polarity detection using machine translation on multilanguage datasets.

In our experiments, we did not have this issue since we already knew the words *a priori*. This result shows that the classic word-by-word translation should not be underestimated regarding sentiment analysis.

We check all the 501k outputs for all English sentiment analysis methods running on each translated sentence, and we observe that in only 38k cases the sentiment analysis methods flipped the polarity for the same sentence. This means that only 7% of the output sentences from the translators have inverted polarities. This conclusion does not mean that the sentences are keeping their sentiment polarity from the original language, but it gives confidence that choosing a machine translation system does not make a sentiment analysis method generate contradictory classification.

It is 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. To help understand this trade-off we show the results for both metrics in the next subsection.

From now on, all Macro-F1 and Applicability scores discussed in this work are the majoring votings between the 3 commercial 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. Performance evaluation of machine translation when used for sentiment analysis in multilingual text compared to native methods

In this section, we explain the results after evaluating the sentiment analysis using machine translation to classify sentiments in multiple languages. First, we present in Fig. 6, the distribution of Macro-F1 scores for non-Native methods on each language dataset. To complement this Figure, we have at Appendix A, Tables A.1–A.14, where we show the results for Applicability, F1-scores for positive and negative classes, and Macro-F1 for each language dataset. Additionally, we have Figs. 7–9, where we can visualize the trade-off between Applicability and Macro-F1. Now, we discuss the main findings regarding these results.

In Fig. 6, we want to share one key finding. If you hide the labels on the x-axis, it is very hard to tell accurately which bar corresponds to the English language. This factor indicates that, although the datasets were created under different circumstances as discussed in Section 3, a potential lack of efficiency of the machine translation approach does not seem to influence the general performance of the sentiment analysis methods. If the contrary happens, we would expect the corresponding English boxplot to be higher, 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, also observed by [18]. 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 in machine-translated text. Next, we look into the separated results for each sentiment analysis method in each language dataset, considering the Macro-F1 and Applicability.

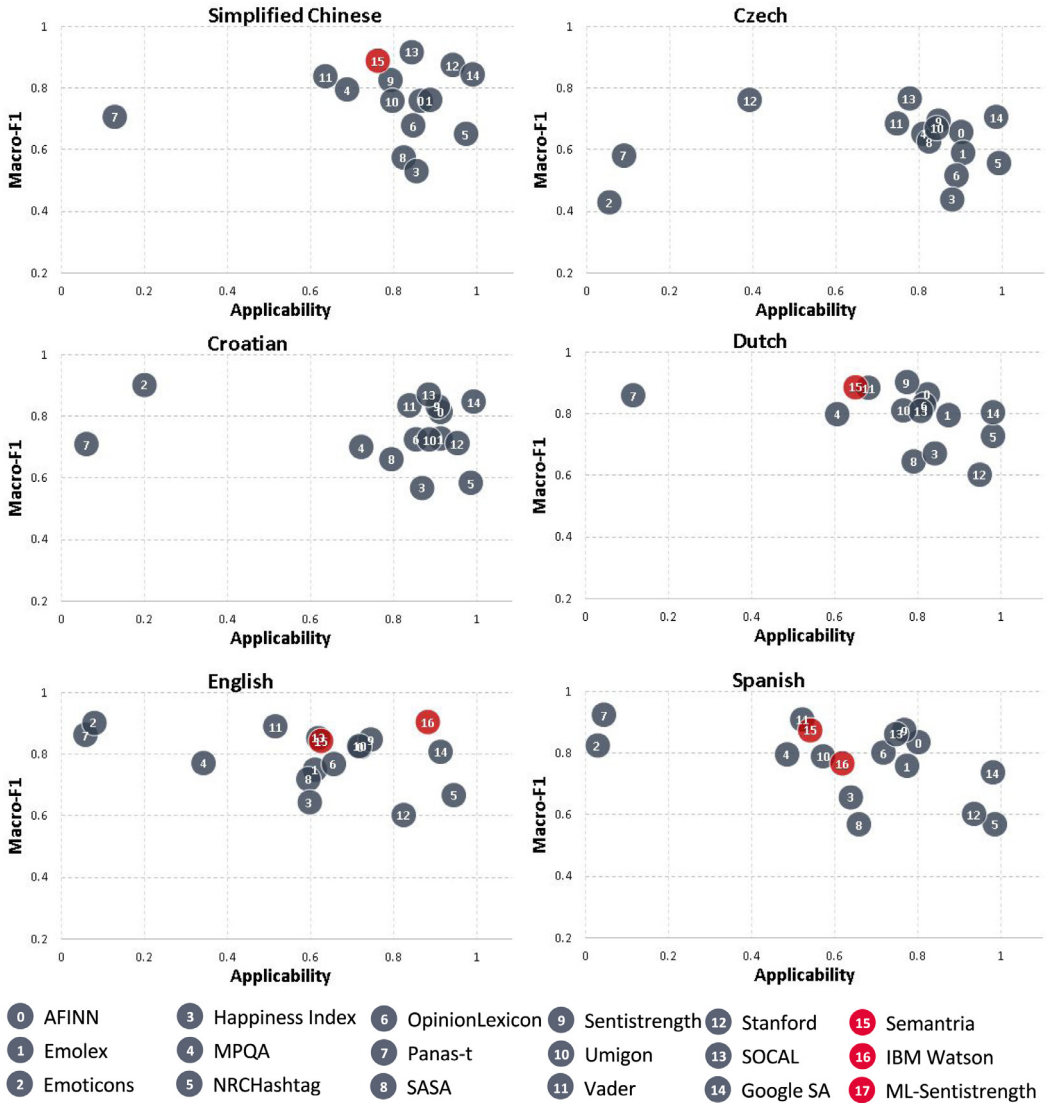

Fig. 7. Macro-F1 vs Applicability.

In Figs. 7–9 we can visualise the trade-off between Macro-F1 and Applicability. 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). As closer to the upper-right corner of the chart, the 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 does not support the corresponding language.

By looking at the charts in Figs. 7–9, we can see that Emoticons(2) appears in the upper-left positions, demonstrating good Macro-F1, but poor Applicability in most languages. For Chinese dataset, SOCAL(13) and Stanford(12) as good methods, with both Macro-F1 and Applicability above 0.8. In the Portuguese dataset, only VADER(11) and Emoticons(2) have Macro-F1 above 0.8, but Vader has a higher Applicability. Overall, Google Sentiment Analysis API(14) highlights as a good approach, presenting a very high Applicability and Macro-F1 often above 0.8, thus, appearing on the right side of almost all the charts. As discussed in Section 4.2, the Haitian Creole chart has the most heterogeneous shape, with many of the methods positioned close to the bottom-left corner.

For instance, regarding the performance of the native methods, we can highlight the IBM Watson(16) for English in Fig. 8, with an outstanding performance in Applicability and Macro-F1. But IBM Watson performs poorly for French dataset, appearing on the bottom-left corner. 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 modest performance, always ranging between 0.6 and 0.8 for both Applicability and Macro-F1.

Fig. 8. Macro-F1 vs Applicability.

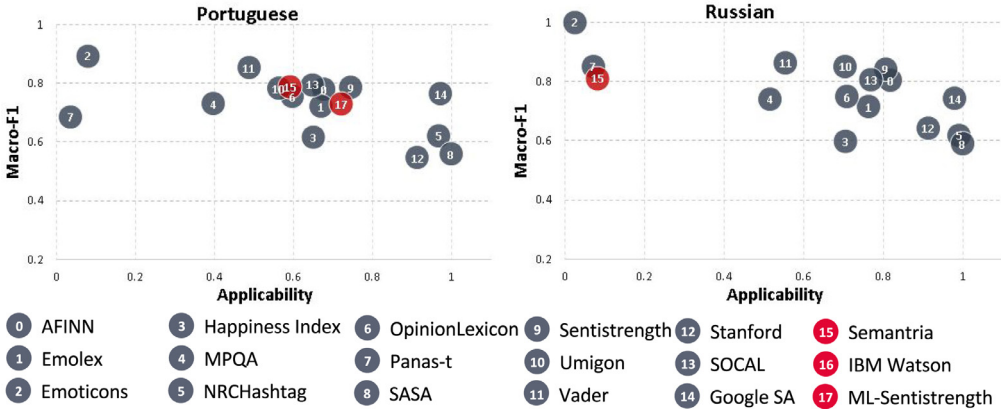

Fig. 9. Macro-F1 vs Applicability.

Table 9
Mean Macro-F1 and Applicability metrics comparing the machine translation approach and native methods.

Language	Macro-F1-Translated	Applicability -Translated	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

The visual findings discussed also manifest in the detailed experiment data presented at Tables from A.1 to Table A.14. In these tables, we can identify a strong variation of the prediction performance 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 sentences as neutral (98%) in the Russian dataset because the lack of emoticons. This emoticon dependency leads the method to a bad performance regarding Applicability for most of the datasets.

Since Tables from A.1 to Table A.14 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 ($F1\text{-}Score_{neg} = 0.86$), 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 performed very well in positive sentences. But, it was poor classifying negative sentences correctly. We noticed 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 Tables A.1–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, SOCAL, Vader, and Google Sentiment Analysis API. This suggests that these methods might be the most reliable ones for multilingual sentiment analysis using machine translation.

As our main goal is to evaluate if machine translation-based methods can perform sentiment analysis as well as the natives methods, we summarize the results, separating both groups of methods (translated and natives). In Table 9 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 on average, but a lower Applicability. However, we need to discuss a caveat of this finding. In the Russian dataset, 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 13 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 consistently 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 if our hypothesis holds.

5. 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 to perform multilingual sentiment analysis. Moreover, we grouped the results from each approach 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.

To solve this problem, we sort the methods by each evaluation metric in each language dataset. Then, we summarize our results in a table, where, in one column we show the average position across our languages datasets with its confidence interval ($\alpha = 0.95$), and in another column, we show the average score of the chosen metric. In Table 10 we show these results considering the Macro-F1, and in Table 11 we show the results for Applicability.

In Table 10, 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 below Emoticons and

Table 10
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 11
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

Vader, but its average position appears at 5th in the rankings. After Semantria, the best native method is IBM Watson, 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. Next, we evaluate the average position performance based on the Applicability metric.

Examining Table 11, where we rank the methods according to Applicability, we have interesting findings. First, the Google Sentiment Analysis API and NRCHashtag appear in the 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, 10 of our 13 methods that use the translation technique show better results than the best native method (ML-Sentistrength).

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 Tables A.1–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 holds, thus, English state-of-the-art sentiment analysis methods combined with machine translator systems can be as good, or even better than "off-the-shelf" native methods. This result triggers an alert for authors of native methods, showing that they should compare their new methods not only with other native methods but also with the machine translation and further analysis of state-of-the-art English methods.

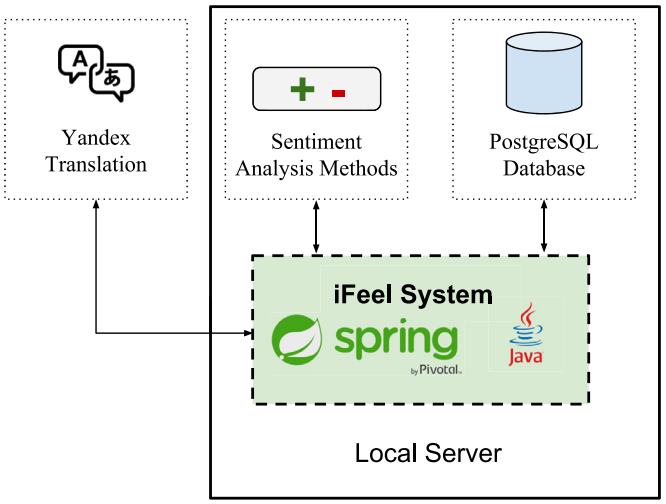

Fig. 10. iFeel 3.0 Architeture.

6. iFeel system

We presented how methods developed for English text, with the help of machine translators, can be better than methods engineered specifically for a non-English language. Thus, we want to make not just this methodology available, but also the whole set of methods easily accessible for others in the scientific community.

Aiming reproducibility, we propose iFeel 3.0,[13] a benchmark system for sentence-level multilingual sentiment analysis. First published in [2], iFeel implemented only eight methods without multilingual support. On its second version published at [1] we increased the set of methods to 14 and also introduced the multilingual approach presented in this work. Despite both previous publications of the system and the high acceptance from the scientific community, we decided to rebuild iFeel, now in its third version.

Re-developing iFeel was motivated by the lack of scalability and stability in both previous versions. The system had a high peak of 100 users, and due to its high computational resources demands, the system crashed when few users upload files to be analyzed in parallel. Additionally, iFeel 2.0 was develop using the Meteor Framework[14], a NodeJs based framework for fast development and prototyping. However, because of Meteor constant changes, updates and libraries deprecated the manage of iFeel 2.0 code was unsustainable. Thus, we chose to recreate iFeel from scratch using the Spring Framework[15] and Java as the backend programming language. Spring is stable and meant to support ”in production” applications.

6.1. iFeel architecture and functionalities

The architecture of iFeel is represented in Fig. 10. The local server runs iFeel and is responsible for storing data, provide a security layer and respond to user requests while they are interacting with the system. When 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. A PostgreSQL database is responsible for saving the sentences uploaded and also data from registered users. Finally, to perform multilingual sentiment analysis iFeel uses the Yandex Translate API with the same methodology presented in this work. Yandex was chosen because it has the largest free tier among the top commercial machine translator systems.

In the index page, we allow the user to test the system with one sentence and receive the response of all sentiment analysis methods implemented. We leave two fields to be filled by the user, the language option, and a free text field to perform the sentiment analysis as shown in Fig. 11. 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 Stanford and Happiness Index classified as “neutral”. After the users register themselves, they have access to the “Analyse File Texts” page, where we allow the upload of multiple sentences from a text file. The upload page is shown in Fig. 12. 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 sentiment analysis for each line of the file with a maximum of 1000 sentences. The result is a *.xml* or *.xlsx* file which the user can download containing the output of all methods implemented.

[13] iFeel is hosted on http://www.ifeel.dcc.ufmg.br, iFeel-resources (code and datasets) is available at https://homepages.dcc.ufmg.br/~fabricio/ifeel_resources.htm.

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

[15] https://spring.io/.

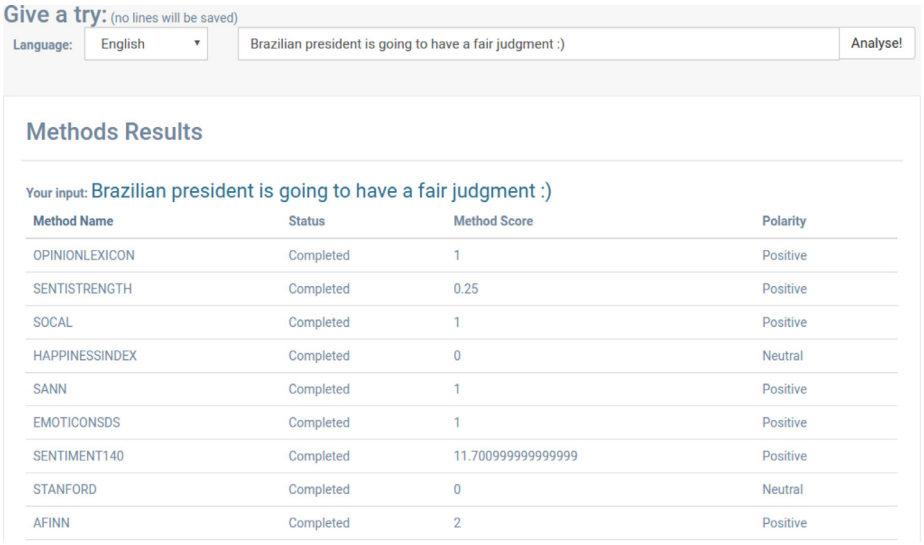

Fig. 11. iFeel - First user experience.

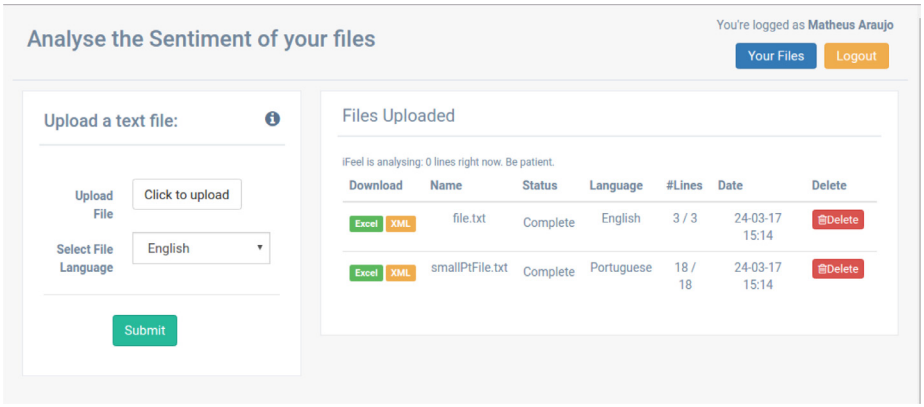

Fig. 12. iFeel - File upload section.

A future step for iFeel is to provide a REST API for its users. The ability to use iFeel automatically as an API is by far the most requested functionality by our users. It meets the need of the current state of the Web 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.

7. Conclusion

The Sentiment analysis field is currently popular and important for 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 academic and commercial application. However, it is still limited by English-only targets, not only off-the-shelf tools but also methodologies of how to solve the problem. Therefore, 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.

The sentiment analysis methods created for the English language perform as well as a specific-language method when the text is translated to English using machine translation. To support this statement we present the results for English and native methods throughout all datasets in different languages, 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 verified which approach is better, comparing the average Macro-F1 and Applicability across datasets. Then, using the average position across the languages datasets to rank 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. SOCAL or Sentistrength showed to be the better option to combine with machine translation for non-English texts as input.

We also 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 translator tools. Our conclusion regarding this topic is that machine translators are stable, showing consistent results among them all. But more important, when we compared the commercial translators with our baseline, there was not a big difference. Thus, the classic word-by-word translation should not be underestimated regarding sentiment analysis. In fact, word-by-word translation could be desired if one has a large enough translation dictionary and concern about processing time and translation cost.

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

We believe in two main direct applications 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 instead of developing a solution for 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 developed the new iFeel 3.0 system using Spring/Java framework which provides a more stable and reliable sentiment analysis environment. 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[16].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by individual grants from CNPq, CAPES, and Fapemig.

Appendix A

Table A1
Simplified Chinese.

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

[16] The datasets are available at https://homepages.dcc.ufmg.br/~fabricio/ifeel_resources.htm.

Table A2
German.

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 A3
Spanish.

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 A4
Greek.

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 A5
French.

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 A6
Croatian.

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 A7
Hindi.

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 A8
Dutch.

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 A9
Czech.

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 A10
Haitian Creole.

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 A11
English.

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 A12
Portuguese.

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 A13
Russian.

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 A14
Italian.

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

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ins.2019.10.031.

References

[1] M. Araújo, J.P. Diniz, L. Bastos, E. Soares, M. Ferreira, F. Ribeiro, F. Benevenuto, iFeel 2.0: a multilingual benchmarking system for Sentence-level sentiment analysis, in: Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17–20, 2016., 2016, pp. 758–759, doi:10.1145/2851613.2851817.

[2] M. Araújo, P. Gonçalves, M. Cha, F. Benevenuto, ifeel: a system that compares and combines sentiment analysis methods, in: Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 75–78, doi:10.1145/2567948.2577013.

[3] P. Arora, Sentiment Analysis for Hindi Language, International Institute of Information Technology Hyderabad, 2013 Ph.d. thesis.

[4] B.W. Bader, W.P. Kegelmeyer, P.A. Chew, Multilingual sentiment analysis using latent semantic indexing and machine learning, in: 2011 IEEE 11th International Conference on Data Mining Workshops, 2011, pp. 45–52, doi:10.1109/ICDMW.2011.185.

[5] A. Balahur, M. Turchi, Multilingual sentiment analysis using machine translation? in: Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis, Association for Computational Linguistics, 2012, pp. 52–60.

[6] C. Banea, R. Mihalcea, J. Wiebe, S. Hassan, Multilingual subjectivity analysis using machine translation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 127–135, doi:10.3115/1613715.1613734.

[7] P. Basile, N. Novielli, Uniba at evalita 2014-sentipolc task: predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features, Proceedings of EVALITA (2014) 58–63, doi:10.18653/v1/s15-2099.

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

[9] P.S. Dodds, C.M. Danforth, Measuring the happiness of large-scale written expression: songs, blogs, and presidents, J. Happiness Stud. 11 (2009) 441–456, doi:10.1007/s10902-009-9150-9.

[10] G. Glavaš, D. Korenčić, J. Šnajder, Aspect-oriented Opinion Mining from User Reviews in Croatian, in: Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 18–23. URL http://www.aclweb.org/anthology/W13-2404.

[11] P. Gonçalves, M. Araújo, F. Benevenuto, M. Cha, Comparing and combining sentiment analysis methods, in: Proceedings of the first ACM conference on online social networks, ACM, 2013, pp. 27–38, doi:10.1145/2512938.2512951.

[12] P. Goncalves, F. Benevenuto, M. Cha, PANAS-t: A pychometric scale for measuring sentiments on twitter, 2013, URL http://arxiv.org/abs/1308.1857v1 abs/1308.1857v1.

[13] M. Hu, B. Liu, Mining and summarizing customer reviews, in: KDD '04, 2004, pp. 168–177, doi:10.1145/1014052.1014073.

[14] C. Hutto, E. Gilbert, Vader: a parsimonious rule-based model for sentiment analysis of social media text, in: Eighth International AAAI Conference on Weblogs and Social Media, 2014, pp. 216–225.

[15] T. Kincl, M. Novák, J. Přibil, Getting inside the minds of the customers: automated sentiment analysis, in: European Conference on Management Leadership and Governance ECMLG, 2013, pp. 122–129.

[16] O.Y. Koltsova, S. Alexeeva, S. Kolcov, An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media, in: Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2016 (Moscow), 2016, pp. 277–287.

[17] C. Levallois, 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), Association for Computational Linguistics, Atlanta, Georgia, USA, 2013, pp. 414–417. URL http://www.aclweb.org/anthology/S13-2068.

[18] W. Lewis, Haitian creole: how to build and ship an mt engine from scratch in 4 days, 17 h, & 30 min, in: EAMT 2010: Proceedings of the 14th Annual conference of the European Association for Machine Translation, European Association for Machine Translation, 2010.

[19] S.L. Lo, E. Cambria, R. Chiong, D. Cornforth, Multilingual sentiment analysis: from formal to informal and scarce resource languages, Artif. Intell. Rev. (2016) 1–29, doi:10.1007/s10462-016-9508-4.

[20] B. Lu, C. Tan, C. Cardie, B.K. Tsou, 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, in: HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 320–330. URL http://dl.acm.org/citation.cfm?id=2002472.2002514.

[21] N. Makrynioti, V. Vassalos, Sentiment extraction from tweets: Multilingual challenges, in: S. Madria, T. Hara (Eds.), Big Data Analytics and Knowledge Discovery: 17th International Conference, DaWaK 2015, Valencia, Spain, September 1–4, 2015, Proceedings, Springer International Publishing, Cham, 2015, pp. 136–148, doi:10.1007/978-3-319-22729-0_11.

[22] P.F. Melo, D.H. Dalip, M.M. Junior, M.A. Gonçalves, F. Benevenuto, 10Sent: a stable sentiment analysis method based on the combination of off-the-shelf approaches, JASIST 70 (2019) 242–255, doi:10.1002/asi.24117.
[23] X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu, H. Wang, Cross-lingual mixture model for sentiment classification, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, 2012, pp. 572–581.
[24] R. Mihalcea, C. Banea, J. Wiebe, Learning multilingual subjective language via cross-lingual projections, in: Proceedings of the 45th annual meeting of the association of computational linguistics, 2007, pp. 976–983.
[25] S. Mohammad, #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), Association for Computational Linguistics, Montréal, Canada, 2012, pp. 246–255. URL http://www.aclweb.org/anthology/S12-1033.
[26] S. Mohammad, P.D. Turney, Crowdsourcing a word-emotion association lexicon, Comput. Intell. 29 (2013) 436–465.
[27] S. Narr, M. Hulfenhaus, S. Albayrak, Language-independent twitter sentiment analysis, Knowl. Discov. Mach. Learn. (KDML), LWA (2012) 12–14.
[28] F.A. Nielsen, A new anew: Evaluation of a word list for sentiment analysis in microblogs, 2011, URL http://arxiv.org/abs/1103.2903.
[29] R. Plutchik, A general psychoevolutionary theory of emotion, in: Emotion: Theory, research, and experience: Vol. 1. Theories of emotion, Academic press, New York, 1980, pp. 3–33.
[30] E. Refaee, V. Rieser, Benchmarking machine translated sentiment analysis for arabic tweets, in: HLT-NAACL, 2015, pp. 71–78, doi:10.3115/v1/n15-2010.
[31] J. Reis, F. Benevenuto, P. Vaz de Melo, R. Prates, H. Kwak, J. An, Breaking the news: first impressions matter on online news, in: Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, Oxford, UK, May 26–29, AAAI Press, 2015, pp. 357–366.
[32] F.N. Ribeiro, M. Araújo, P. Gonçalves, M.A. Gonçalves, F. Benevenuto, Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods, EPJ Data Sci. 5 (2016) 1–29, doi:10.1140/epjds/s13688-016-0085-1.
[33] S. Ruder, P. Ghaffari, J.G. Breslin, INSIGHT-1 at semeval-2016 task 5: deep learning for multilingual aspect-based sentiment analysis, CoRR abs/1609.02748 (2016), doi:10.18653/v1/s16-1053.
[34] A.A. Rios, P.J. Amarilla, G.A.G. Lugo, Sentiment categorization on a creole language with lexicon-based and machine learning techniques, in: 2014 Brazilian Conference on Intelligent Systems, 2014, pp. 37–43, doi:10.1109/BRACIS.2014.18.
[35] G. Shalunts, G. Backfried, N. Commeignes, The impact of machine translation on sentiment analysis, Data Anal. 2016 (2016) 63.
[36] K. Shuang, Z. Zhang, H. Guo, J. Loo, A sentiment information collector–extractor architecture based neural network for sentiment analysis, Inf. Sci. 467 (2018) 549–558, doi:10.1016/j.ins.2018.08.026.
[37] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts, Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank, in: 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1631–1642.
[38] M. Souza, R. Vieira, Sentiment analysis on twitter data for portuguese language, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 241–247, doi:10.1007/978-3-642-28885-2_28.
[39] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2011) 267–307, doi:10.1162/COLI_a_00049.
[40] M. Thelwall, Heart and soul: Sentiment strength detection in the social web with sentistrength, 2013, http://sentistrength.wlv.ac.uk/documentation/SentiStrengthChapter.pdf. doi: 10.1007/978-3-319-43639-5_7,
[41] E. Tromp, Multilingual Sentiment Analysis on Social Media, Lap Lambert Academic Publ, 2012.
[42] Q. V. Le, M. Schuster, A neural network for machine translation, at production scale, 2016, URL https://research.googleblog.com/2016/09/a-neural-network-for-machine.html.
[43] J. Villena Román, S. Lana Serrano, E. Martínez Cámara, J.C. González Cristóbal, TASS-workshop sentiment analysis at SEPLN, Procesamiento del Lenguaje Natural 50 (2013) 37–44.
[44] X. Wan, Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, in: EMNLP '08, Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, pp. 553–561.
[45] H. Wang, D. Can, A. Kazemzadeh, F. Bar, S. Narayanan, A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle, in: The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea, The Association for Computer Linguistics, 2012, pp. 115–120.
[46] D. Watson, L. Clark, Development and validation of brief measures of positive and negative affect: the panas scales, J. Pers. Soc. Psychol. 54 (1985) 1063–1070, doi:10.1037/0022-3514.54.6.1063.
[47] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan, Opinionfinder: a system for subjectivity analysis, in: HLT/EMNLP on Interactive Demonstrations, 2005, pp. 34–35, doi:10.3115/1225733.1225751.
[48] C.-S. Yang, H.P. Shih, A rule-based approach for effective sentiment analysis, in: PACIS Proceedings, 2012, p. 181.
[49] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: a survey, Wiley Interdisci. Revi. Data Mining Knowl. Discov. 8 (2018) e1253, doi:10.1002/widm.1253.
[50] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, B. Xu, Text classification improved by integrating bidirectional lstm with two-dimensional max pooling, in: Proceedings of the 26th international conference on computational linguistics (COLING 2016), 2016, pp. 3485–3495, doi:10.1145/2851613.2851817. https://www.aclweb.org/anthology/C16-1329.