Comparative Analyses of Multilingual Sentiment Analysis Systems for News and Social Media

Pavel Přibáň[1,2] and Alexandra Balahur[1]

[1] European Commission Joint Research Centre
Via E. Fermi 2749, 21027 Ispra (VA), Italy
[2] University of West Bohemia
Faculty of Applied Sciences, Department of Computer Science and Engineering
Univerzitni 8, 301 00 Plzeň, Czech Republic
pribanp@kiv.zcu.cz, alexandra.balahur@ec.europa.eu

Abstract. In this paper, we present evaluation of three in-house sentiment analysis (SA) systems originally designed for three distinct SA tasks, in a highly multilingual setting. For the evaluation, we collected a large number of available gold standard datasets, in different languages and varied text types. The aim of using different domain datasets was to achieve a clear snapshot of the level of overall performance of the systems and thus obtain a better quality of an evaluation. We compare the results obtained with the best performing systems evaluated on their basis and performed an in-depth error analysis. Based on the results, we can see that some systems perform better for different datasets and tasks than the ones they were designed for, showing that we could replace one system with another and gain an improvement in performance. Our results are hardly comparable with the original dataset results because the datasets often contain a different number of polarity classes than we used, and for some datasets, there are even no basic results. For the cases in which a comparison was possible, our results show that systems perform well in view of multilinguality.

1 Introduction

Recent years have seen a growing interest in the task of Sentiment Analysis (SA). In spite of these efforts however, real applications for sentiment analysis are still challenged by a series of aspects, such as multilinguality and domain dependence. Sentiment analysis can be divided into different sub-tasks like aspect based SA, polarity or fine-grained SA, entity-centered SA. SA can also be applied on many different levels of scope – document level, sentence or phrases level. Performing sentiment analysis in a multilingual setting is even more challenging, as most datasets available are annotated for English texts and low-resourced languages suffer from a lack of annotated datasets on which machine learning models can be trained.

In this paper, we describe an evaluation of our three in-house SA systems designed for three distinct SA tasks, in a highly multilingual setting. These systems process a tremendous amount of text every day, and therefore it is essential to know their quality and also be able to evaluate these applications correctly. At present, these systems cannot be sufficiently evaluated. Due to the lack of correct evaluation, we decided to prepare appropriate resources and tools for the evaluation, assess these applications and summarize obtained results. We collect and describe a rich collection of publicly available datasets for sentiment analysis, and we present the performance of individual systems for the collected datasets. We also carry out additional experiments with the datasets, and we show that for news articles performance of classification increases when adding the title of the news article to the body text.

1.1 Tasks Description

The evaluated systems are intended for solving three sentiment related tasks – **Twitter Sentiment Analysis** (*TSA*) task, **Tonality in News** (*TON*) task and the **Targeted Sentiment Analysis** (*ESA*) task that can also be called Entity-Centered Sentiment Analysis.

In the **Twitter Sentiment Analysis** and **Tonality** tasks, the systems have to assign a polarity which determines the overall sentiment of a given tweet or a news article (generally speaking text).

Targeted Sentiment Analysis (*ESA*) task is a task of a sentiment polarity classification towards an entity mention in a given text.

For all mentioned tasks, the sentiment polarity can be one of the *positive, negative* or *neutral* labels or a number from -100 to 100, where a negative value indicates negative sentiment, a positive value indicates positive sentiment and zero (or values close to zero) means neutral sentiment. In our evaluation experiments, we used the 3-point scale (*positive, negative, neutral*).

1.2 Systems Overview

`TwitOMedia` system [3] for the *TSA* task uses a hybrid approach, which employs supervised learning with a Support Vector Machines Sequential Minimal Optimization [32], on unigram and bigram features.

`EMMTonality` system for the *TON* task counts occurrences of language specific sentiment terms from our in-house language specific dictionaries. Each sentiment term has a sentiment value assigned. The system sums up values for all words (which are present in the mentioned dictionary) in a given text. The resulting number is normalized and scaled to a range from -100 to 100 where the negative value indicates negative tonality, the positive value indicates positive tonality and the neutral tonality is expressed with zero.

`EMMTonality` system also contains a module for the *ESA* task which computes sentiment towards an entity in a given text. This approach is the same as for the tonality in news articles, with the difference that only a certain number of words surroundings the entity are used to compute the sentiment value towards the entity.

EMMSenti system is intended to solve only the *ESA* task. This system uses a similar approach to the EMMTonality system, see [38] for the detailed description.

2 Related Work

In [35], authors summarize eight publicly available datasets for a Twitter sentiment analysis and they are giving an overview of the existing evaluation datasets and their characteristics. Another comparison of available methods for sentiment analysis is mentioned in [14]. They describe four different approaches (*machine learning, lexicon-based, statistical* and *rule-based*) and they distinguish between three different levels of the scope of sentiment analysis, i.e. *document level*, *sentence level* and *word/phrase/sub-sentence level*.

In recent years most of the state-of-the-art systems and approaches for sentiment analysis used neural networks and deep learning techniques. Very popular became the Convolutional Neural Network (CNN) [24] and the Recurrent Neural Network (RNN) like Long Short-Term Memory (LSTM) [21] or Gated Recurrent Unit (GRU) [11]. Kim [22] used a CNN architecture for sentiment analysis and question answering. One of the proofs of neural networks successfulness is that most of the top teams [7, 13, 18] in sentiment analysis (or tasks related to the sentiment analysis) in the last SemEval [34, 29] and WASSA [23, 27] competitions used deep learning techniques. Zhang et al. [41] present a comprehensive survey of current application in sentiment analysis. Barnes et al. [4] compare several models on six different benchmark datasets, which belong to different domains and additionally have different levels of granularity. They showed that LSTMs based neural networks are particularly good at fine-grained sentiment tasks. Tang et al. [39] introduced sentiment-specific word embeddings (SSWE) for Twitter sentiment classification, which encode sentiment information in the continuous representation of words.

The majority of the sentiment analysis research mainly focuses on monolingual methods, especially in English but some effort is being made for multilingual approaches as well. Balahur and Turchi [1] propose an approach to obtain training data for French, German and Spanish using three distinct Machine translation (MT) systems. They translated English data to the three languages, and then they evaluated performance for sentiment analysis after using the three MT systems. They showed that the gap in classification performance between systems trained on English and translated data is minimal, and they claim that MT systems are mature enough to be reliably employed to obtain training data for languages other than English and that sentiment analysis systems can obtain comparable performances to the one obtained for English. In [2] they extended work from [1] and showed that tf-idf weighting with unigram features has a positive impact on the results.

In [10], the authors study possibilities of usage of English model for sentiment analysis in different Russian, Spanish, Turkish and Dutch languages where the annotated data are more limited. They propose a multilingual approach where a single RNN model is built in the language where the largest sentiment analysis resources are available. Then they used MT to translate test data to English and finally they used the model to classify the translated data.

Dashtipour et al. [15] provide a review of multilingual sentiment analysis. They compare their implementation of existing approaches on common data. Precision observed in their experiments is typically lower than the one reported by the original authors, which could be caused by the lack of detail in the original presentation of those approaches.

Zhou et al. [42] created bilingual sentiment word embeddings, which is based on the idea of encoding sentiment information into semantic word vectors. Related multilingual approach for sentiment analysis for low-resource languages is presented in [6]. They introduced Bilingual Sentiment Embeddings (BLSE), which are jointly optimized to represent (a) semantic information in the source and target languages, which are bound to each other through a small bilingual dictionary, and (b) sentiment information, which is annotated on the source language only.

In [5], authors extend an approach from [6] to domain adaption for sentiment analysis. Their model takes as input two mono-domain embedding spaces and learns to project them to a bi-domain space, which is jointly optimized to project across domains and to predict sentiment.

From the previous review, we can deduce that the current state-of-the-art approaches for sentiment analysis in English are solely based on neural networks and deep learning techniques. Deep learning techniques usually require more data than the ”traditional” machine learning approaches (Support Vector Machine, Logistic Regression) and it is evident that they will be used for rich-resources languages (English). On the other hand, much less effort was invested in the multilingual approaches, and low-resources languages compared to English. First studies about multilingual approaches mostly relied on machine translation systems, but in recent years neural networks along with deep learning techniques were employed as well. Another common idea for multilingual approaches in SA is that researchers are trying to find a way how to create a model based on data from rich-resources language and transform the knowledge in such a way that it is possible to use the model for other languages.

3 Datasets

In this section, we describe collected datasets for the evaluation. The evaluated applications require different types of datasets or at least different domains to carry out a proper evaluation. We collected mostly public available datasets, but we also used our in-house non-public datasets. The polarity labels for all collected twitter and news datasets are *positive, neutral or negative.* If the original dataset contained other polarity labels than the three mentioned we discard them or mapped them to *positive, neutral* or *negative* polarity labels.

Sentiment analysis of tweets is a prevalent problem, and much effort is being put into solving this problem and related problems in recent years [19, 34, 20, 30, 28, 27, 23]. Therefore, datasets for this task are easier to find.

On the other hand, finding datasets for the *ESA* task is much more challenging because there is less research effort being put into this task and thus there are less existing resources. For the sentiment analysis in news articles we were not able to find a proper public dataset for the English language, and therefore we used our in-house datasets. For some languages exist publicly available corpora such as Slovenian [9], German [25], Brazilian Portuguese [16], Ukrainian and Russian [8].

3.1 Twitter Datasets

In this subsection, we present the sentiment datasets for the Twitter domain. We collected 2.8M labelled tweets in total from several datasets, see Table 1 for detailed statistics. Next, we shortly describe each of these datasets.

Table 1: Twitter datasets statistics.

Dataset	**Total**	**Positive**	**Negative**	**Neutral**
Sentiment140 Test	498	182	177	139
Sentiment140 Train	1 600 000	800 000	800 000	-
Health Care Reform	2394	543	1381	470
Obama-McCain Debate	1904	709	1195	-
Sanders	3424	519	572	2333
T4SA	1 179 957	371 341	179 050	629 566
SemEval 2017 Train	52 806	20 555	8430	23 821
SemEval 2017 Test	12 284	2375	3972	5937
InHouse Tweets Test	3813	1572	601	1640
InHouse Tweets Train	4569	2446	955	1168
Total	2 861 649	1 200 242	996 333	665 074

Sentiment140 [19] dataset consists of two parts – training and testing. The training part includes 800k positive and 800k negative automatically labelled tweets. Authors of this dataset collected tweets containing certain emoticons, and to every tweet, they assigned a label based on the emoticon. For example, :) and :-) both express positive emotion and thus tweets containing these emoticons were labelled as positive. The testing part of this dataset is composed of 459 manually annotated tweets (177 negative, 177 neutral and 182 positives). The detailed description of this approach is described in [19].

Speriosu et al. [37] created **Health Care Reform** dataset based on tweets about health care reform in the USA. They extracted tweets containing the health care reform hashtag "#hcr" from the early 2010s. This dataset contains 543 positive, 1381 negative and 470 neutral examples.

Obama-McCain Debate [36] dataset was manually annotated with the *Amazon Mechanical Turk* by one or more annotators for the categories *positive,* *negative, mixed* or *other*. Total 3269 tweets posted during the presidential debate on September 26th, 2008 between Barack Obama and John McCain were annotated. We filtered this dataset to obtain only tweets with a *positive* or *negative* label (no *neutral* classes were present). After the filtering process, we received 709 positives and 1195 negative examples.

T4SA [40] dataset was obtained from July to December 2016. The authors discarded retweets, tweets not containing any static image and tweets whose text was less than five words long. Authors were able to gather 3.4M tweets in English. Then, they classified the sentiment polarity of the texts and selected the tweets having the most confident textual sentiment predictions. This approach resulted in approximately a million labelled tweets. For the sentiment polarity classification, authors used an adapted version of the ItaliaNLP Sentiment Polarity Classifier [12]. This classifier uses a tandem LSTM-SVM architecture. Along with the tweets, authors also crawled the images contained in the tweets. The aim was to automatically build a training set for learning a visual classifier able to discover the sentiment polarity of a given image [40].

SemEval-2017 dataset was created for the *Sentiment Analysis in Twitter* task [34] at SemEval 2017. The authors made available all the data from previous years of the Sentiment Analysis in Twitter [30] tasks and they also collected some new tweets. They chose English topics based on popular events that were trending on Twitter. The topics included a range of named entities (e.g., *Donald Trump, iPhone*), geopolitical entities (*e.g., Aleppo, Palestine*), and other entities. The dataset is divided into two parts – *SemEval 2017 Train* and *SemEval 2017 Test*. They used CrowdFlower to annotate the new tweets.

We removed all duplicated tweets from the *SemEval 2017 Train* part which resulted in approximately 20K positive, 8K negative and 23K neutral examples and 2K positive, 4K negative and 6K neutral examples for the *SemEval 2017 Test* part (see Table 1).

InHouse Tweets dataset consists of two datasets *InHouse Tweets Train* and *InHouse Tweets Test* used in [3]. These datasets come from SemEval 2013 task 2 *Sentiment Analysis in Twitter* [20].

Sanders twitter dataset[3] created by *Sanders Analytics* consists of 5512 manually labelled tweets by one annotator. Each tweet is related to one of four topics (Apple, Google, Microsoft, Twitter). Tweets are labelled as either *positive, negative, neutral* or *irrelevant*. We discarded tweets labelled as *irrelevant*. Saif et al. [35] also described and used *Sanders* twitter dataset.

3.2 Targeted Entity Sentiment Datasets

For the *ESA* task, we were able to collect three labelled datasets. Datasets from [17, 26] are created from tweets, and our *InHouse Entity* dataset [38] contains sentences from news articles. Detailed statistics are shown in Table 2.

[3] Dataset can be obtained from https://github.com/pmbaumgartner/text-feat-lib

Table 2: Targeted Entity Sentiment Analysis datasets statistics.

Dataset	**Total**	**Positive**	**Negative**	**Neutral**
Dong	6940	1734	1733	3473
Mitchel	3288	707	275	2306
InHouse Entity	1281	169	189	923
Total	11 509	2610	2197	6702

Dong [17] is manually annotated dataset for the *ESA* task consisting of 1734 positive, 1733 negative and 3473 neutral examples. Each example consists of a tweet, an entity and a class label which denotes a sentiment towards the entity.

Mitchell et al. [26] used the *Amazon Mechanical Turk* to annotate **Mitchel** dataset with 3288 examples (tweet – entity pairs) for the *ESA* task. Tweets with a single highlighted named entity were shown to the annotators, and they were instructed to select the sentiment being expressed towards the entity (*positive*, *negative* or *no sentiment*).

For the evaluation, we also used our **InHouse Entity** dataset created by Steinberger et al. [38]. This dataset was created as a multilingual parallel news corpus annotated with sentiment towards entities. They used data from Workshops on Statistical Machine Translation (2008, 2009, 2010)[4]. Firstly, they recognized the named entities and then selected examples were manually annotated with two annotators. The disagreed cases were judged by the third annotator. They were able to obtain 1281 labelled examples (707 positive, 275 negative and 923 neutral), e.g. sentences with annotated entity and sentiment expressed towards the entity.

3.3 News Tonality Datasets

For the *TON*[5] task, we used our two non-public multilingual datasets. Firstly, our **InHouse News** dataset consists of 1830 manually labelled texts from news articles about Macedonian Referendum in 23 languages, but the majority is formed by Macedonian, Bulgarian, English, Italian and Russian, see Table 3. Each example contains a title and description of a given article. For the evaluation of our systems we used only Bulgarian, English, Italian and Russian because other languages are either not supported by the evaluated systems or the number of examples is less than 60 samples.

EP News dataset contains more than 50K manually labelled news articles about the European Parliament and European Union in 25 European languages. Each news article in this dataset consists from a title and full text of the article and also from their English translation, we selected five main European languages

[4] http://www.statmt.org/wmt10/translation-task.html

[5] For this task we also used tweets described in subsection 3.1

(English, German, French, Italian and Spanish) for the evaluation, see Table 4 for details.

Table 3: InHouse News dataset statistics.

InHouse News	**Total**	**Positive**	**Negative**	**Neutral**
Macedonian	974	516	234	224
Bulgarian	215	118	26	71
English	339	198	35	106
Italian	62	41	3	18
Russian	65	17	34	14
Other Languages	175	60	44	71
Total	1830	950	376	504

Table 4: EP Tonality News dataset statistics.

EP News	**Total**	**Positive**	**Negative**	**Neutral**
English	2193	263	172	1758
German	5122	389	179	4554
French	2964	574	308	2082
Italian	1544	291	152	1101
Spanish	3594	324	135	3135
Total	15417	1841	946	12630

4 Evaluation & Results

In this section, we present the summary of all the evaluation results for of all the three systems. For each system, we select an appropriate collection of datasets, and we classify examples of each selected dataset separately. Then, we merge all selected datasets, and we classify them together. Except for the *InHouse News* dataset and *EP News* dataset, all experiments are performed on English texts. We carry out experiments on the `EMMTonality` system with the *InHouse News* dataset on Bulgarian, English, Italian and Russian. Experiments with the *EP News* dataset are performed on the `TwitOMedia` and `EMMTonality` system with English, German, French, Italian and Spanish[6].

[6] On the `EMMTonality` system we perfom experiments with all available languages, but we report results only for English, German, French, Italian and Spanish.

Each sample is classified as `positive, negative` or `neutral` and for all named systems we did not apply any additional preprocessing steps[7]. As an evaluation metric, we used `Accuracy` and `Macro` F_1 score which are defined as:

$$F_1^M = \frac{2 \times P^M \times R^M}{P^M + R^M} \tag{1}$$

where P^M denotes `Macro Precision` an R^M denotes `Macro Recall`. Precision P_i and recall R_i are firstly computed separately for each class (n is the number of classes) and then averaged as follows:

$$P^M = \frac{\sum_i^n P_i}{n} \tag{2}$$

$$R^M = \frac{\sum_i^n R_i}{n} \tag{3}$$

4.1 Baselines

For basic comparison, we created baseline models for the *TSA* task and *TON* task. These baseline models are based on unigram or unigram-bigram features. Results are shown in tables 5, 6, 7, 8 and 9. For the baseline models, we apply minimal preprocessing steps like lowercasing and word normalization which includes conversion of URLs, emails, money, phone numbers, usernames, dates and numbers expressions to one common token, for example, token "www.google.com" is converted to the token "<url>". These steps lead to a reduction of feature space as shown in [19]. We use `ekphrasis` library from [7] for word normalization.

Table 5: Results of baseline models for the InHouse Tweets Test dataset with **unigram features** (models were trained on InHouse Tweets Train dataset).

Baseline	**Macro F_1**	**Accuracy**
Log. regression	**0.5525**	**0.5843**
SVM	0.5308	0.5641
Naive Bayes	0.4233	0.4993

To train the baseline models, we use an implementation of Support Vector Machines (SVM) – concretely Support Vector Classification (SVC) with `linear kernel`, Logistic Regression with `lbfgs` solver and Naive Bayes algorithms from the `scikit-learn` library [31], default values are used for other parameters of the mentioned classifiers. Our *InHouse News* dataset does not contain a large number of examples, and therefore we perform experiments with 10-fold cross-validation, the same approach is applied for the *EP News* dataset.

[7] Except baselines systems.

Table 6: Macro F_1 score and Accuracy results of baseline models with **unigram** and **bigram** features. The `InHouse News` dataset and the `EP News` dataset with all examples (**all languages**) were used. We used 10-fold cross-validation (results in table are averages of individual folds). Bold values denote best results for each dataset.

		InHouse News		**EP News**	
Baseline	**Config**	**Macro F_1**	**Accuracy**	**Macro F_1**	**Accuracy**
Log. regression	text	0.663	0.705	0.551	0.864
	text+title	0.704	0.738	0.578	**0.870**
SVM	text	0.657	0.697	0.564	0.856
	text+title	**0.717**	**0.747**	**0.591**	0.866
Naive Bayes	text	0.612	0.676	0.513	0.845
	text+title	0.646	0.702	0.552	0.852

For the News datasets (*InHouse News* and *EP News*) we train baseline models with different combinations of data. In Table 6 are shown results for models which are trained on a concatenation of examples in different languages. For each dataset, we select all untranslated examples (texts in original languages), and we train model regardless of the language. The model is then able to classify texts in all languages which were used to train the model. This approach should lead to performance improvement as is shown in [3]. The same approach is used to acquire results for Table 7, but only specific languages are used, specifically for the *InHouse News* dataset it is English, Bulgarian, Italian and Russian and for the *EP News* dataset it is English, French, Italian, German and Spanish. Table 8 contains results for models trained only on original English texts. In tables 6, 7, 8 and 9 column `Config` denotes whether the text of an example is used or if a title of the example is concatenated with the text and is used as well.

If we compare baseline results from Table 8 with results from Table 10 (last five lines of the table), we can see that baselines perform much better than our current system (see Macro F_1 score in tables). The `TwitOMedia` system was trained on tweets messages, so it is evident that its performance on news articles will be lower, but the `EMMTonality system` should achieve better results.

Our results from tables 6, 7 and 8 confirm the claims from [3] that joining of data in different languages leads the performance improvement. Models trained on all examples (regardless language), see Table 6, achieve best results.

We collected large manually labelled dataset of tweets, and we wanted to study the possibility to use this dataset to train a model. This model would then be used for classification of news articles that are different from the domain of the training data. After comparing results from Table 9 (last five lines of the table) with results from Table 10, we can see that our simple baseline is not outperformed on the *InHouse News* dataset by the other two systems. These

Table 7: Macro F_1 score and Accuracy results of baseline models with **unigram** and **bigram** features. The *InHouse News* dataset with **Bulgarian, English, Italian** and **Russian** examples and the *EP News* dataset with **English, French, Italian, German** and **Spanish** examples were used. We used 10-fold cross-validation (results in table are averages of individual folds). Bold values denote best results for each dataset.

		InHouse News		**EP News**	
Baseline	**Config**	**Macro F_1**	**Accuracy**	**Macro F_1**	**Accuracy**
Log. regression	text	0.629	0.682	0.497	0.833
	text + title	**0.692**	**0.729**	0.529	**0.841**
SVM	text	0.630	0.677	0.513	0.819
	text + title	0.684	0.718	**0.540**	0.833
Naive Bayes	text	0.585	0.657	0.432	0.816
	text + title	0.612	0.678	0.457	0.817

Table 8: Macro F_1 score and Accuracy results of baseline models with **unigram** and **bigram** features. The *InHouse News* dataset and *EP News* dataset only with original **English** examples were used. We used 10-fold cross-validation (results in table are averages of individual folds). Bold values denote best results for each dataset.

		InHouse News		**EP News**	
Baseline	**Config**	**Macro F_1**	**Accuracy**	**Macro F_1**	**Accuracy**
Log. regression	text	0.612	0.730	0.510	0.820
	text + title	**0.685**	**0.769**	0.534	0.826
SVM	text	0.608	0.719	0.530	0.815
	text + title	0.674	0.760	**0.546**	**0.827**
Naive Bayes	text	0.502	0.695	0.441	0.818
	text + title	0.547	0.713	0.446	0.819

results show that it is possible to use data from different domains for training and obtain good results.

We also observe that incorporating the title (concatenating the title and the text) of a news article leads to an increase in performance across all datasets and combination of data used for training models. These results show that the title is an essential part of the news and contains significant sentiment and semantic information despite its short length.

4.2 Twitter Sentiment Analysis

To evaluate a system for the *TSA* task, we used a domain rich collection of tweets datasets. We collected datasets with almost 3M labelled tweets, detailed

Table 9: Macro F_1 score and Accuracy results of baseline models trained on *SemEval 2017 Train* and *Test* datasets with **unigram** features. Evaluation was performed on original English examples from our *InHouse News* and *EP News* datasets. Bold values denote best results for each dataset.

		InHouse News		**EP News**	
Baseline	**Config**	**Macro F_1**	**Accuracy**	**Macro F_1**	**Accuracy**
Log. regression	text	0.395	0.432	0.312	0.518
	text + title	**0.408**	**0.462**	0.310	0.495
SVM	text	0.380	0.429	0.283	0.408
	text + title	0.389	0.456	0.287	0.397
Naive Bayes	text	0.237	0.296	0.313	**0.639**
	text + title	0.239	0.293	**0.314**	0.620

statistics of used datasets can be seen in Table 1. Table 10 shows obtained results for *Accuracy* and *Macro F_1* measures.

From Table 10 is evident that the `TwitOMedia` system [3] performs best for the *InHouse Tweets Test* dataset (bold values in the table). This dataset is based on data from [20] and was used to develop (train and test) this system.

The reason why the `TwitOMedia` system performs better for the *InHouse Tweets Test* dataset than for the *InHouse Tweets Train* dataset (HTTr) is that the system was trained on translations of the HTTr dataset. Original training dataset (HTTr) was translated into several languages, and then the translations were merged to one training dataset which was used to train the model. This approach leads to performance improvement as is shown in [3].

For the other datasets, the performance is lower especially for the domain-specific ones and datasets which does not contain instances with *neutral* classes, for example, *Health Care Reform* dataset or *Sentiment 140 Train* dataset. The first reason is most likely that the system was trained on the other domain of texts which is too much different and thus the system is not able to successfully classify (generalize) texts from these domain-specific datasets. Secondly, *Sentiment140 Train* dataset and *Obama-McCain Debate* dataset do not contain examples with a *neutral* class.

4.3 Tonality in News

`EMMTonality` system for the *TON* task was evaluated on the same set of datasets like the one for the `TwitOMedia` system. Obtained results are shown in table 10.

If we compare results of the `TwitOMedia` system and results of the `EMMTonality` system, we can see that the `EMMTonality` system achieves better results for these datasets: *Sentiment140 Test, Health Care Reform, Obama-McCain Debate, Sanders, SemEval 2017 Train*, and *SemEval 2017 Test*. The overall results are better for the `TwitOMedia` system. Results for the *InHouse News* and *EP News* datasets are comparable for both evaluated systems.

Table 10: Macro F_1 score and Accuracy results of the evaluated `TwitOMedia` and `EMMTonality` systems. Bold values denote the best results in specific dataset category (Individual Twitter datasets, joined Twitter datasets and News datasets), and underlined values denote best results for specific dataset category and for each system separetely.

	TwitOMedia		**EMMTonality**	
Dataset	**Macro F_1**	**Accuracy**	**Macro F_1**	**Accuracy**
Sentiment140 Test	0.566	0.530	0.666	0.639
Health Care Reform	0.410	0.326	0.456	0.403
Obama-McCain Debate (OMD)	0.270	0.290	0.331	0.357
Sanders	0.468	0.591	0.526	0.618
Sentiment140 Train (S140T)	0.312	0.358	0.250	0.375
SemEval 2017 Train	0.501	0.529	0.538	0.561
SemEval 2017 Test	0.460	0.500	0.552	0.564
T4SA	0.603	0.669	0.410	0.392
InHouse Tweets Test (HTT)	**0.710**	**0.708**	0.583	0.610
InHouse Tweets Train (HTTr)	0.629	0.599	0.580	0.574
All Tweets w/o S140T, OMD, T4SA	**0.597**	**0.660**	0.545	0.563
All Tweets w/o S140T, T4SA	0.507	0.528	0.542	0.558
InHouse News en	0.397	0.425	0.398	0.425
EP News en, text	0.368	**0.698**	0.422	0.678
EP News en, title + text	0.372	0.690	**0.425**	0.675
EP News translated, text	0.368	0.432	0.390	0.278
EP News translated, title + text	0.369	0.388	0.393	0.238

Regarding multilinguality, the `EMMTonality` system slightly overperforms the `TwitOMedia` system in Macro F_1 score, see Table 11. Table 11 contains results for the *EP News* dataset for five European languages (English, German, French, Italian and Spanish).

4.4 Targeted Sentiment Analysis

We evaluated the `EMMSenti` system and `EMMTonality` system for the *ESA* task on the *Dong, Mitchel* and *InHouse Entity* datasets, see table 12 for results.

We obtained the best results for the *InHouse Entity* dataset in terms of *Accuracy* measure and also for the *Macro* F_1 score. The best results across all datasets and systems are obtained for the *neutral* class (not reported in the table) and for other classes our systems work more poorly. The classification algorithm (for both systems) is based on counting subjective terms (words) around entity mentions (no machine learning algorithm or approach is involved). It is obvious that the quality of dictionaries used, as well as their adaptation to the domain, is crucial. If no subjective term from the text is found in the dictionary, to the example is assigned the neutral label.

Table 11: Macro F_1 score and Accuracy results for the *EP News* dataset for English, German, French, Italian and Spanish examples.

		TwitOMedia		**EMMTonality**	
Lang.	**Config**	**Macro F_1**	**Accuracy**	**Macro F_1**	**Accuracy**
EN	Text	0.368	0.698	0.422	0.678
	Text+Title	0.372	0.690	0.425	0.675
DE	Text	0.333	0.711	0.348	0.846
	Text+Title	0.344	0.687	0.360	0.730
FR	Text	0.354	0.614	0.389	0.549
	Text+Title	0.356	0.602	0.383	0.472
IT	Text	0.314	0.692	0.397	0.347
	Text+Title	0.351	0.690	0.405	0.330
ES	Text	0.337	0.828	0.392	0.386
	Text+Title	0.332	0.823	0.392	0.333

The best performance of our systems for the neutral class can be explained by the fact that most of the neutral instances do not contain any subjective term.

We also have to note that we were not able to reproduce results obtained in [38] and our achieved performance for this dataset is worse. It is possible that Steinberger et al. [38] used slightly different lexicons than we used.

Table 12: Macro F_1 score and Accuracy results for the `EMMSenti` and `EMMTonality` systems evaluation. Bold values denote best results for each dataset.

	EMMSenti		**EMMTonality**	
Dataset	**Macro F_1**	**Accuracy**	**Macro F_1**	**Accuracy**
Dong	0.491	0.512	0.496	0.501
Mitchel	0.483	0.660	0.490	0.640
InHouse Entity	**0.517**	**0.663**	0.507	0.659
All	0.505	0.571	0.512	0.557

4.5 Error Analysis

In order to understand the causes leading in erroneous classification, we analyze the misclassified examples from Twitter and the News datasets for the `EMMTonality` and the `TwitOMedia` systems. We categorize the errors into four groups (see below)[8]. We randomly select 40 incorrectly classified examples for each class and for each system across all datasets which were used for evaluation of these systems, which resulted in 240 manually evaluated examples in total.

We found the following major groups of errors:

1. Implicit sentiment/external knowledge: Sentiment is often expressed implicitly, or external knowledge is needed for a correct classification. The evaluated text does not contain any explicit attributes (words, phrases, emoji/emoticons) which would clearly indicate the sentiment and because our systems are based on surface level features (unigrams/bigrams or counting occurrences of sentiment words), they will fail in these examples. For example, text like *"We went to Stanford University today. Got a tour. Made me want to go back to college."* indicates positive sentiment but for this decision we have to know that *Stanford University* is prestigious university (which is positive) and according to the sentence *"Made me want to go back to college."* author has probably a positive relation to universities or his previous studies. This group of errors is the most common in our set of the error analysis examples, we observed it in 94 cases and only for positive or negative examples.

2. Slang expression: Misclassified examples in this group contain domain-specific words, slang expressions, emojis, unconventional linguistic means, misspelled or uppercased words like *"4life", "YEAH BOII", "yessss", "grrrl", "yummmmmy"*. We observe this type of errors in 29 examples and most of them were caused by the `EMMTonality` system (which is reasonable because this system is intended for news). The appropriate solution for part of this problem is an application of preprocessing steps like spell correction, lowercasing, text normalization (*"yesssss"* ⇒ *"yes"*) or extending of dictionaries. In case of extending dictionaries, we have to deal with Twitter vocabulary because the Twitter vocabulary (vocabulary of tweets) is changing quite fast (new modern expressions and hashtags are introduced often) and thus dictionaries have to be modified regularly. On the other hand, the `TwitOMedia` system would have to be retrained every time with new examples in order to extend its feature set or a more advanced normalization system should be used in the pre-processing stage.

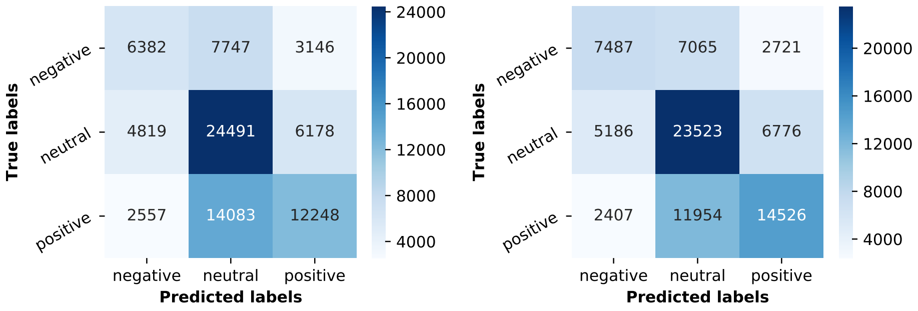
3. Negation: Negation of terms is an essential aspect of sentiment classification [33]. Negations can easily change or reverse the sentimental orientation. This error appeared in 35 cases in our set of the error analysis examples.

4. Opposite sentiment words: The last type of errors is caused by sentiment words which express the opposite or different sentiment than the entire text. This type of error was typical for examples annotated with a `neutral` label. For example tweet *"#Yezidi #Peshmerga forces playing volleyball and crushing #ISIS in the frontline."* is annotated as neutral but contains words like *"crushing, #ISIS"* or *"frontline"* which can indicate negative sentiment. We observed this error in 20 examples.

[8] Each incorrectly classified example may be contained in more than one error group. Some examples were also (in our view) annotated incorrectly. For some cases, we were not able to discover the reason for misclassification.

The first group of errors (`Implicit sentiment/external knowledge`) was the most common among the evaluated examples and is also the hardest one because the system would have to have access to world knowledge or be able to detect implicit sentiment in order to be able of correct classification. This error was observed only for examples annotated with positive or negative labels; there, the explicit sentiment markers are missing. The majority of these examples were misclassified as a neutral class. In this case, the sentiment analysis system must be complemented with a system for emotion detection similar to one of the top systems from [23] to improve classification performance. In case of emotion detection for examples which were classified as a neutral class, we would change the neutral class according to the detected emotion. The examples with negative emotions like sadness, fear or anger would be changed to the negative class and examples with positive emotions like joy or surprise would be changed to the positive class.

Figure 1 shows confusion matrices for the `EMMTonality` and `TwitOMedia` systems. We can see that a noticeable amount of misclassified examples was predicted as a neutral class and the mentioned improvement should positively affect a significant number of examples according to our statistics from the error analysis.



(a) TwitOMedia system

(b) EMMTonality system

Fig. 1: Confusion matrices for the `TwitOMedia` and `EMMTonality` systems on all tweets without *S140T* and *T4SA* datasets.

Lastly, we have to note that we were not able to decide the reason for misclassification in 35 cases. According to us, in seven cases was the annotated label incorrect.

5 Conclusions

In this paper, we showed the process of thoroughly evaluating three systems for sentiment analysis with a comparison of their performance. We collected and described a rich collection of publicly available datasets, and we performed experiments with these datasets and showed the performance of individual systems. We carried out additional experiments with collected datasets and showed that for news articles is more beneficial also to include the title of the news article along with the text of the article itself. We performed a thorough error analysis and proposed potential solutions for each category of misclassified examples.

In our future work, we will explore current state-of-the-art methods and develop new approaches (including deep learning methods, multilingual embeddings and other recent machine learning approaches) for multilingual sentiment analysis in order to implement them in our highly multilingual environment.

Acknowledgments

This work was partially supported from ERDF "Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)" (no.: CZ.02.1.01/0.0/0.0/17_048/0007267) and by Grant No. SGS-2019-018 Processing of heterogeneous data and its specialized applications.

Bibliography

[1] Alexandra Balahur and Marco Turchi. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 52–60, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2392963.2392976.

[2] Alexandra Balahur and Marco Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75, 2014.

[3] Alexandra Balahur, Marco Turchi, Ralf Steinberger, Jose Manuel Perea Ortega, Guillaume Jacquet, Dilek Küçük, Vanni Zavarella, and Adil El Ghali. Resource creation and evaluation for multilingual sentiment analysis in social media texts. In *LREC*, pages 4265–4269. Citeseer, 2014.

[4] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–12. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-5202. URL http://aclweb.org/anthology/W17-5202.

[5] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. Projecting embeddings for domain adaption: Joint modeling of sentiment analysis in diverse domains. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 818–830. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/C18-1070.

[6] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2483–2493. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/P18-1231.

[7] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[8] Victoria Bobichev, Olga Kanishcheva, and Olga Cherednichenko. Sentiment analysis in the ukrainian and russian news. In *Electrical and Computer Engineering (UKRCON), 2017 IEEE First Ukraine Conference on*, pages 1050–1055. IEEE, 2017.

[9] Jože Bučar, Martin Žnidaršič, and Janez Povh. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52(3):895–919, 2018.

[10] Ethem F. Can, Aysu Ezen-Can, and Fazli Can. Multilingual sentiment analysis: An rnn-based framework for limited data. *CoRR*, abs/1806.04511, 2018.

[11] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179. URL http://aclweb.org/anthology/D14-1179.

[12] Andrea Cimino and Felice Dell'Orletta. Tandem lstm-svm approach for sentiment analysis. In *CLiC-it/EVALITA*, 2016.

[13] Mathieu Cliche. Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 573–580. Association for Computational Linguistics, 2017. doi: 10.18653/v1/S17-2094. URL http://aclweb.org/anthology/S17-2094.

[14] Anaıs Collomb, Crina Costea, Damien Joyeux, Omar Hasan, and Lionel Brunie. A study and comparison of sentiment analysis methods for reputation evaluation. *Rapport de recherche RR-LIRIS-2014-002*, 2014.

[15] Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771, 2016.

[16] Gabriel Domingos de Arruda, Norton Trevisan Roman, and Ana Maria Monteiro. An annotated corpus for sentiment analysis in political news. In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, pages 101–110, 2015.

[17] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *The 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 2014.

[18] Venkatesh Duppada, Royal Jain, and Sushant Hiray. Seernet at semeval-2018 task 1: Domain adaptation for affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 18–23. Association for Computational Linguistics, 2018. doi: 10.18653/v1/S18-1002. URL http://aclweb.org/anthology/S18-1002.

[19] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

[20] J Hltcoe. Semeval-2013 task 2: Sentiment analysis in twitter. *Atlanta, Georgia, USA*, 312, 2013.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[22] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for

Computational Linguistics, 2014. doi: 10.3115/v1/D14-1181. URL http://aclweb.org/anthology/D14-1181.

[23] Roman Klinger, Orphee De Clercq, Saif Mohammad, and Alexandra Balahur. Iest: Wassa-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL http://aclweb.org/anthology/W18-6206.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[25] Andreas Lommatzsch, Florian Bütow, Danuta Ploch, and Sahin Albayrak. Towards the automatic sentiment analysis of german news and forum documents. In *International Conference on Innovations for Community Services*, pages 18–33. Springer, 2017.

[26] Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, 2013.

[27] Saif M. Mohammad and Felipe Bravo-Marquez. WASSA-2017 shared task on emotion intensity. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Copenhagen, Denmark, 2017.

[28] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.

[29] Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

[30] Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Fabrizio Sebastiani. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June 2016. Association for Computational Linguistics.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

[32] John C Platt. 12 fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, pages 185–208, 1999.

[33] Johan Reitan, Jørgen Faret, Björn Gambäck, and Lars Bungum. Negation scope detection for twitter sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 99–108, 2015.

[34] Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[35] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. 2013.

[36] David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*, pages 3–10. ACM, 2009.

[37] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63. Association for Computational Linguistics, 2011.

[38] Josef Steinberger, Polina Lenkova, Mijail Kabadjov, Ralf Steinberger, and Erik Van der Goot. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 770–775, 2011.

[39] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565. Association for Computational Linguistics, 2014. doi: 10.3115/v1/P14-1146. URL http://aclweb.org/anthology/P14-1146.

[40] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. Cross-media learning for image sentiment analysis in the wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 308–317, Oct 2017. doi: 10.1109/ICCVW.2017.45.

[41] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis : A survey. *CoRR*, abs/1801.07883, 2018. URL http://arxiv.org/abs/1801.07883.

[42] HuiWei Zhou, Long Chen, Fulin Shi, and Degen Huang. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 430–440. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-1042. URL http://aclweb.org/anthology/P15-1042.