Multilingual Sentiment Analysis on Data of the Refugee Crisis in Europe

Gayane Shalunts

SAIL LABS Technology GmbH
Vienna, Austria
Email: gayane.shalunts@sail-labs.com

Gerhard Backfried

SAIL LABS Technology GmbH
Vienna, Austria
Email: gerhard.backfried@sail-labs.com

***Abstract*—The refugee crisis in Europe was one of the biggest challenges in summer-autumn 2015. The problem drew the highest attention in media and was the discussion topic of politicians and responsible organizations. The current article presents multilingual sentiment analysis of the traditional media content covering the topic. Sentiment analysis forms an integral part of multifaceted media analysis. The dataset comprises relevant articles from eighty of the most circulated traditional media sources in English, German, Russian and Spanish, compiled in the course of three months. The temporal sentiment classification per language demonstrates how the attitude towards the crisis differs across the languages and geographical areas. The further sentiment analysis and visualizations of various aspects illustrate in details the distribution of positivity/negativity among media sources and their target languages.**

Keywords–Multilingual sentiment analysis; refugee crisis.

I. INTRODUCTION

Sentiment analysis refers to a classification task in Natural Language Processing (NLP) community, the goal of which is commonly to determine the objectivity (objective/subjective) or polarity (positive/negative) of the input data. The main parameters defining the scope of a sentiment analysis approach are the target language, domain and media type (traditional or social media). The most common application is the monitoring of public opinions in marketing (product reviews) and politics (election campaigns). Whereas the research field is active, most publications are limited to the domains of movie and product reviews in English only. Sentiment analysis methods can be divided into two broad categories: machine-learning- and lexicon-based methods [1]. Machine learning methods are implemented as supervised binary (positive/negative) classification approaches, in which classifiers are trained on labeled data [1] [2]. However, the dependence on a labeled dataset is considered a major drawback, since labeling is usually costly, time-intensive and even impossible in some cases. In contrast, lexicon-based methods use a predefined set of patterns (referred to as a sentiment dictionary or lexicon) associating each entry with a specific sentiment score and do not require any labeled training data. Here, the challenge lies in designing an appropriate sentiment lexicon. Lexicon-based methods are tuned towards specific target domains, media types and the respective language style, e.g., formal language on traditional media and colloquial language on social media. A comparison of eight state-of-the-art sentiment analysis methods is performed in [1]. All experiments are carried out using two English datasets of Online Social Networks messages. The methods compared are SentiWordNet [3], SASA [4], PANAS-t [5], Emoticons, SentiStrength [6], LIWC [7], SenticNet [8] and Happiness Index [9]. They report that the examined sentiment analysis methods have different levels of applicability on real-world events and vary widely in their agreement on the predicted polarity. Sentiment analysis of the textual data relevant to disasters/crises aims to provide additional structured information to the responsible organizations for situation analysis in various phases of disaster/crisis management [10].

The current article makes the following contributions: 1) Compiles automatically a corpus of news articles covering the refugee crisis in Europe in a period of a quarter in summer-autumn 2015 (36702 articles in total). The articles originate from 80 of the most circulated traditional media sources in English, German, Russian and Spanish, 2) investigates the temporal development of the data volume per language, 3) performs sentiment analysis per language, 4) detects the sentiment polarity across media sources and their languages and generates visualizations of different aspects. The authors choose to employ the SentiSAIL software tool [11] among numerous existing state-of-the-art sentiment analysis methods to carry out the above remarked experimental setup. SentiSAIL performs multidimensional sentiment analysis in terms of languages, domains and media types. It is integrated into the SAIL LABS Media Mining System (MMS) for Open Source Intelligence [12]. MMS is a state-of-the-art Open-Source-Intelligence system, incorporating speech and text-processing technologies [11]. SentiSAIL performs an important part of MMS automatic multifaceted processing of unstructured textual data. It addresses the content of both traditional and social media in English, German, Russian and Spanish, supports the domains of general news and particularly the coverage of disasters/crises. The performance evaluation of SentiSAIL on a trilingual traditional media corpus, as well as on an English social media dataset for comparison with other state-of-the-art methods, is reported in [11]. The experiments in [11] showed that the performance of SentiSAIL and human annotators are equivalent. SentiSAIL was also used to analyze the social media data in German, concerning the European floods 2013 [13]. The choice of SentiSAIL is motivated by the following advantages of applicability in the current scenario: 1) SentiSAIL supports all the languages mentioned, unlike other sentiment analysis approaches, which handle only a single language content. E.g., SentimentWS [14] and [15] target only German, [16] [17] - Russian, Sentitext [18] and [19] - Spanish. The authors in [20] adapted the English semantic orientation system [21] to Spanish, comparing several alternative approaches. 2) SentiSAIL is adapted to the domain of news articles and especially the coverage of disasters/crises in the traditional media. The authors in [22] also target the

 ISBN: 978-1-61208-510-4

domain of news, limited only to English though.

The paper is organized as follows: Section II clarifies the methodology of the SentiSAIL tool. Section III gives detailed information about the experimental corpora. Section IV presents the experimental setup, performance evaluation and results. And finally, Section V draws conclusions from the work presented.

II. SentiSAIL methodology

SentiSAIL is a multilingual sentiment analysis tool addressing the domain of general news and particularly the coverage of disasters/crises in general news [11]. It employs the algorithm of one of the state-of-the-art sentiment analysis methods SentiStrength [6]. SentiStrength, like [21], is a lexicon-based approach, using lexicons of words associated with scores of positive or negative orientation. SentiSAIL also supports stemming of the lexicon patterns, which is particularly important for the processing of inflective languages, such as Russian or German. The intensification/boosting and negation of the lexicon words, as well as the polarity scoring of phrases and idioms, intend to model the structure and semantics of the language observed. The innovative contribution of SentiSAIL [11] lied in expanding the SentiStrength algorithm into new domains (general and disasters/crises related news), multiple languages (English, German, Russian and Spanish) and to the granularity level of full articles. In the scope of the crises domain were considered both natural and humanitarian crises, like the refugee crises in Europe. The adaptation of SentiSAIL to the crises domain was achieved by means of manual compilation of sentiment terms from relevant texts. Examples of such terms are ”donation”, ”volunteer” (positive terms), ”underfeeding”,”xenophobic” (negative terms).

Whereas SentiStrength is optimized for and evaluated on social media content, SentiSAIL targets both social and traditional media data. The social media features are parameterized and may be disabled during traditional media processing. The SentiStrength and SentiSAIL features are compared in [11] on a self-compiled traditional media corpus, reporting the SentiSAIL performance improvement in English to be slight and considerable in German and Russian. SentiSAIL, like [23], solves a dual classification task by classifying a text into one of the following 4classes: positive, negative, mixed (both positive and negative) or neutral (neither positive, nor negative). The dual classification scheme is motivated, as humans exhibit the ability to experience positive and negative emotions simultaneously [24]. The class of the input text is obtained by taking the following steps: 1) the sentiment on the granularity level of line is determined by obtaining a pair of positive/negative scores by averaging the respective positive/negative scores of the sentiment patterns present in the line. Algorithms other than averaging were also employed in this step without a significant impact on the final classification rate [11]. 2) The sentiment on the granularity level of document is calculated likewise as a pair of positive/negative scores by averaging the pairs of the positive/negative scores of all lines respectively. 3) The final sentiment class of a text is produced by double thresholding of the pair of the positive/negative scores on the granularity level document: classification of the positive and negative classes is straightforward. Documents passing both thresholds are classified into the mixed class, those failing both thresholds are classified as neutral.

III. Data Collection

The multilingual corpus covering the humanitarian crisis of refugees in Europe was collected automatically using the SAIL LABS Feeder for web content a web-crawler aimed at the collection of textual content from feeds and web-pages [12]. The tool can be scheduled to collect traditional media sources on a regular basis. A multilingual corpus (English, German, Russian and Spanish) was compiled to reflect a variety of views in particular geographical regions and formed by cultural differences and political influences. Twenty out of the most circulated traditional media sources per language were chosen in order to obtain equal distribution among languages. The period observed is a quarter from July to September 2015

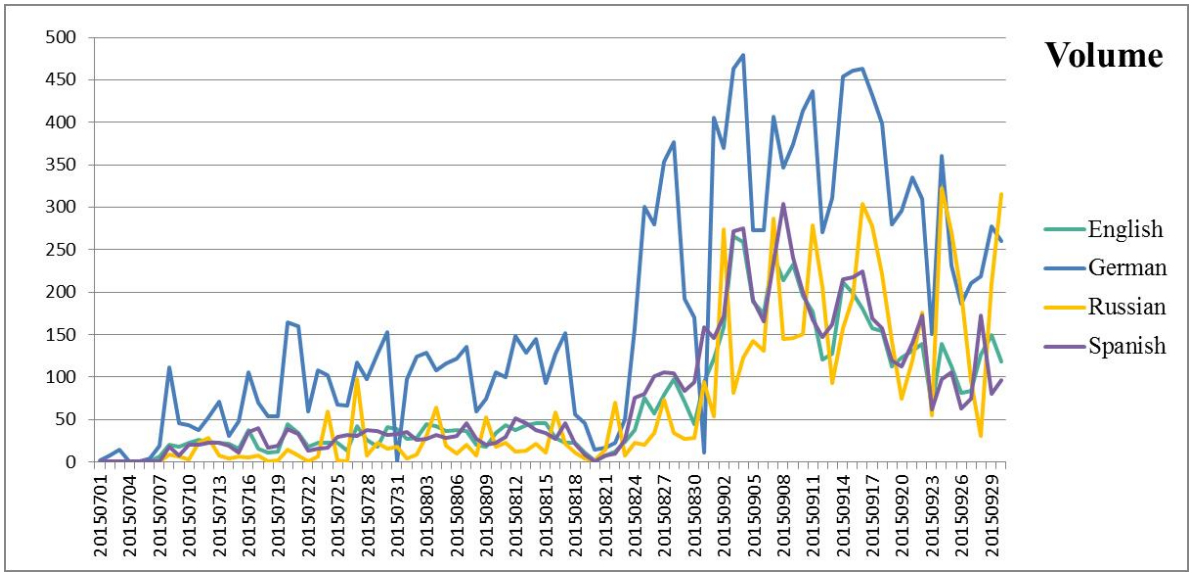

Figure 1. Temporal chart of the data volume per language (horizontal axis - date, vertical axis - article count).

 ISBN: 978-1-61208-510-4

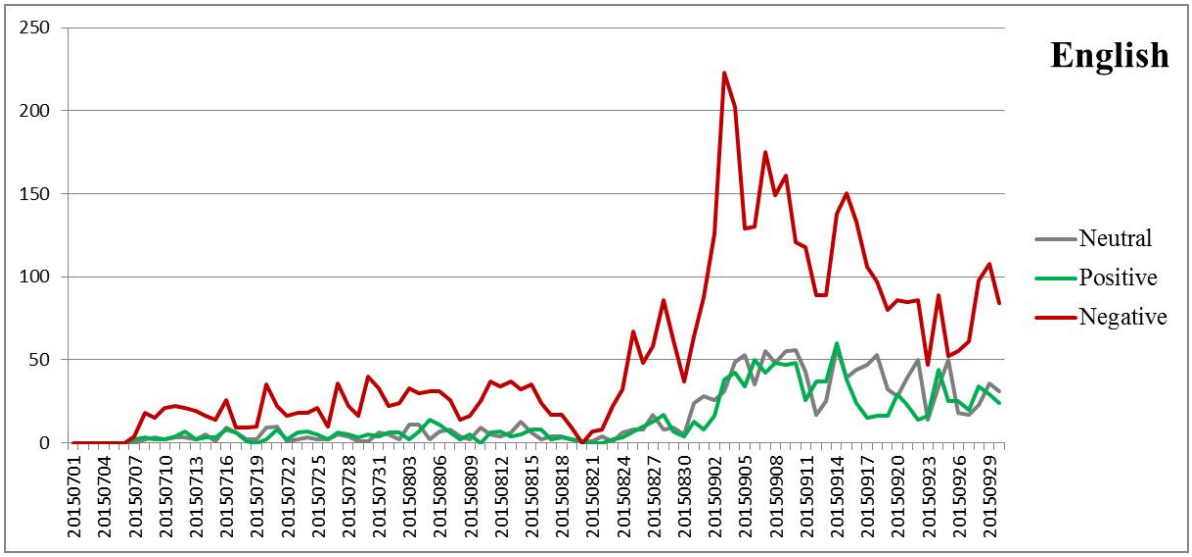

Figure 2. Temporal sentiment analysis of the English data (horizontal axis - date, vertical axis - article count).

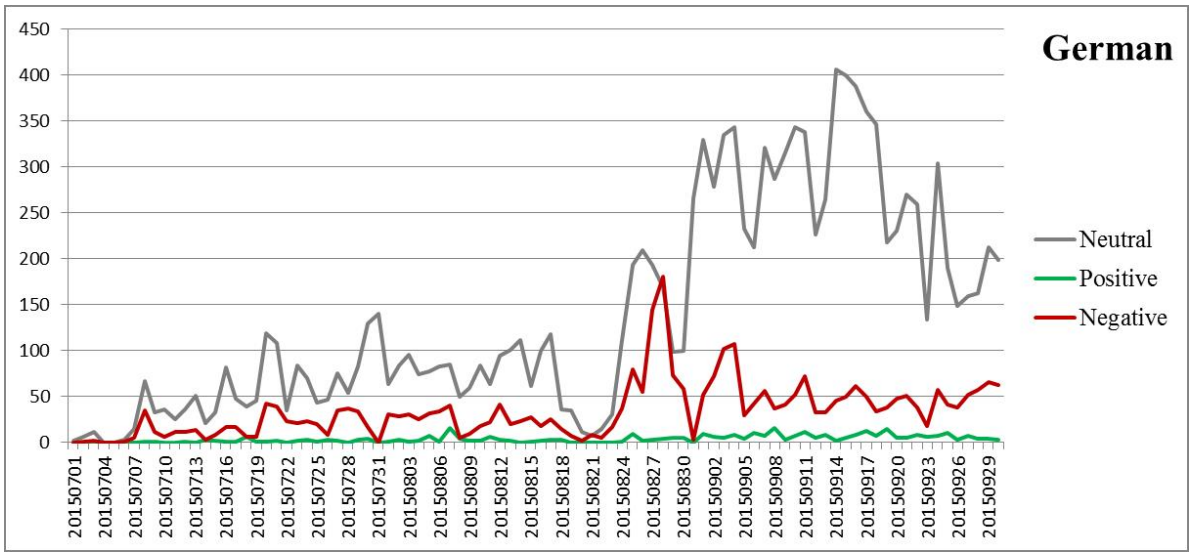

Figure 3. Temporal sentiment analysis of the German data (horizontal axis - date, vertical axis - article count).

on a daily basis. The number of the collected articles in the mentioned period in English is equal to 6580, in German 16669, in Russian 6459 and in Spanish 6994. A total of 36702 articles relevant to the refugee crisis were analyzed.

IV. The Experimental Setup and Results

In order to have a picture about the development of the refugee crisis and how actively the selected multilingual traditional media sources covered those, firstly the temporal distribution of the data volume among languages is examined. Fig. 1 depicts the daily volume of the compiled data per language in the period from July the 1st to September 30th. The data volume growth is noticeable for all languages starting from August 25th, coinciding with the dates of the deepening of the refugee crisis. It is also visible from the chart that the traditional media sources in German paid more attention to the problem, generating the highest volume of content among the four languages monitored. This tendency may be explained by the fact that the German speaking countries Austria and Germany were confronted with the crisis immediately by a vast stream of refugees.

The temporal sentiment analysis of the selected traditional media sources in English, among those CNN and BBC, is displayed in Fig. 2. The vertical axis represents the number of articles per sentiment class, the horizontal axis the issue dates of the articles. The articles, classified in the mixed class, are assigned both to the positive and negative classes. As observed in Fig. 2, the negative sentiment is generally dominating during the whole period. The term refugee(s), associated with a slightly negative score, is excluded from the sentiment patterns in all languages, in order not to bias the classification of the whole corpus towards negative sentiment. Fig. 3, Fig. 4 and

 ISBN: 978-1-61208-510-4

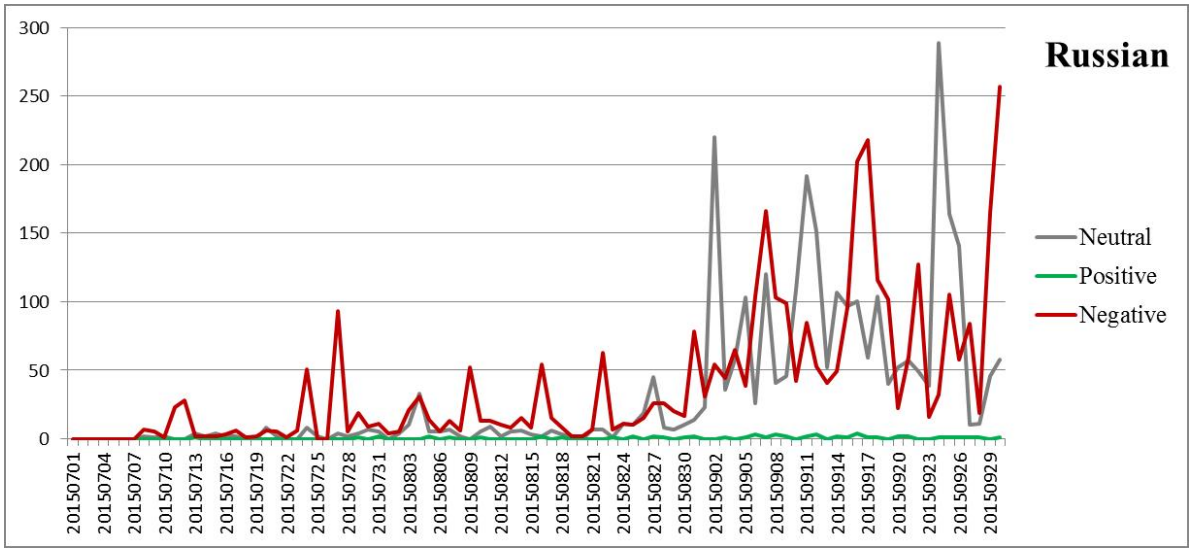

Figure 4. Temporal sentiment analysis of the Russian data (horizontal axis - date, vertical axis - article count).

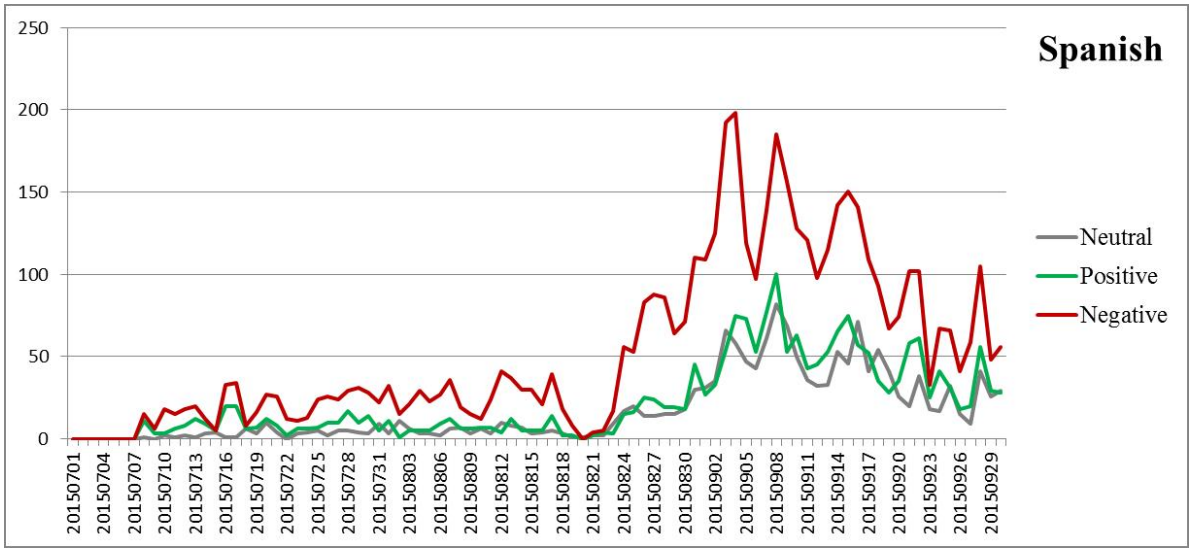

Figure 5. Temporal sentiment analysis of the Spanish data (horizontal axis - date, vertical axis - article count).

Fig. 5 portray the equivalent temporal charts of the distribution of the sentiment classes on German, Russian and Spanish share of the corpus respectively. Comparing the four multilingual charts, one may conclude that whereas the negative sentiment is highly prevailing on English and Spanish media sources, the German media stands out by dominating neutrality. The Russian corpus analysis reveals high rates of both negative and neutral sentiment (Fig. 4). One can also find correlations with the crisis events examining the sentiment distribution diagrams. For example, the global maximum of the negative sentiment on the English data is observed on September 3rd (Fig. 2), coinciding with the date, when a three-year-old boy drowned in his Syrian familys attempt to reach Greece from Turkey. On the other hand, the global negative maximum on the German media is located on August 28th (Fig. 3), when 71 refugees were found dead in the back of a freezer truck in Austria. A closer look at German articles, carrying positive sentiment, revealed phrases, such as Ich bin stolz Deutscher zu sein (I am proud to be German); Eine der größten Spendenaktionen der vergangenen Jahre ist gelungen (One of the biggest donation activities of the past years succeeded); Hilfsbereitschaft (willingness to help), etc.

Table I demonstrates how positivity, negativity and neutrality are distributed among the four languages, covering the sensitive topic of the refugee crisis. The most positive media language is Spanish with 27.61% positive content, the least positive is the Russian one with only 2.35% positive articles. Note for comparison that the average positivity rate of the whole multilingual corpus is 13.13% (Table I). The highest negativity rate is observed on the English media coverage with 72.3% negative rate, whereas the German media spreads the lowest negativity with 23.6% negative rate. The average negativity rate on the whole multilingual corpus yields 52.9%.

 ISBN: 978-1-61208-510-4

TABLE I. THE POSITIVITY, NEGATIVITY AND NEUTRALITY RATES IN PERCENT PER LANGUAGE.

	English	German	Russian	Spanish	Average
Positivity rate %	19.61%	2.96%	2.35%	27.61%	13.13%
Negativity rate %	72.3%	23.6%	45.6%	70.2%	52.9%
Neutrality rate %	19.7%	74.3%	53.2%	21.7%	42.2%

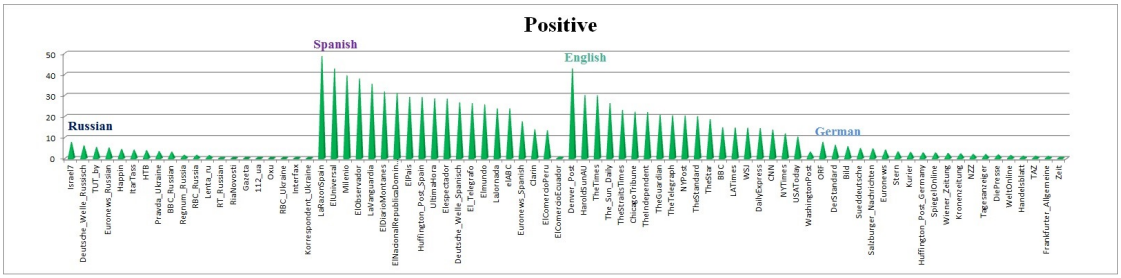

Figure 6. The rate of all positive articles in % per media source.

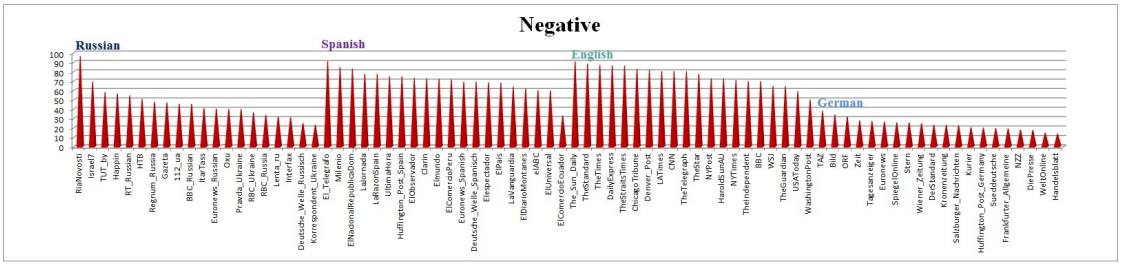

Figure 7. The rate of all negative articles in % per media source.

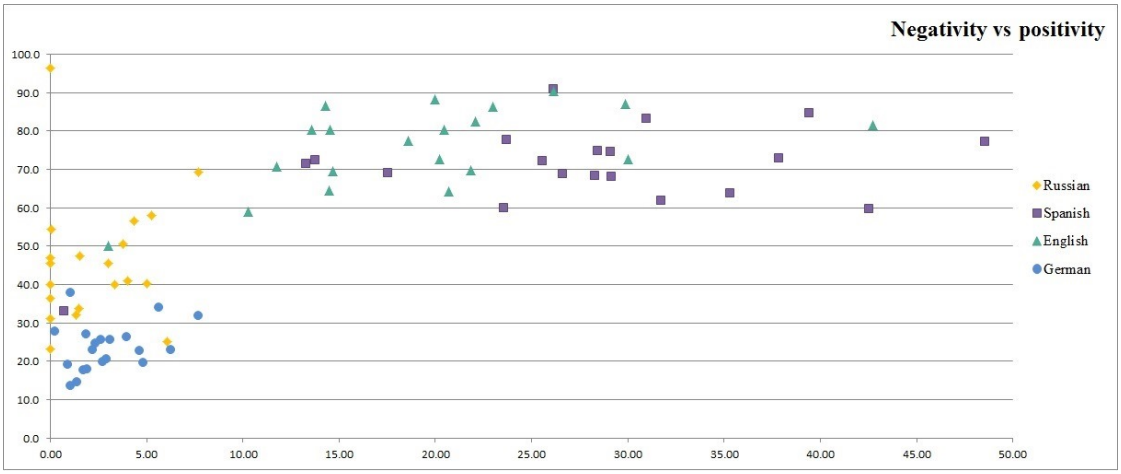

Figure 8. The rate of negativity (vertical axis) vs positivity (horizontal axis) in % per media source and language.

The German corpus also stands out with the highest neutrality rate (74.3%), which is considerably higher than the average neutrality rate of the complete dataset (42.2%) (Table I).

The next facet of our multilingual sentiment analysis portrays the distribution of the positive and negative sentiment per media source to reveal the most positive/negative sources. Fig. 6 depicts the positive rate in percent for the 80 media sources, labeled by their language. Fig. 7 is the corresponding chart for the negative sentiment. The most positive media source is the Spanish La Razon Spain with 48.55% positive content, the least positive ones are the Russian Ria Novosti, Gazeta, 112 Ukraine, Oxu, RBC Ukraine, Interfax and Korrespondent Ukraine, lacking positive content completely (Fig. 6). The most negative media source is the Russian Ria Novosti with 96.3% negative articles, the least negative one - the German Handelsblatt with 13.6% negativity rate (Fig. 7). Fig. 8 shows a visualization of the positivity vs negativity rates in percent per media source and language. Here, an interesting

tendency of clustering per language is noticeable: 1) The German media sources shape a well-defined cluster of low negativity and low positivity. 2) The Russian media sources form a cluster of low positivity and moderate negativity. Here, the clear outlier is Ria Novosti with 96.3% negative and 0% positive content. The clustering is not so clearly notable in cases of the English and Spanish corpora. Here, one may conclude, that whereas both media languages spread highly negative content (exceeding 59%), the range of the distribution of the positivity is very broad. The outliers from the tendency are the English Washington Post (3% positivity vs 50% negativity) and the Spanish El Comercio Ecuador (0.74% positivity vs 33.1% negativity).

V. CONCLUSION

The paper presented sentiment analysis of traditional media data on the 2015 refugee crisis in Europe, originating from a vast number of multilingual, highly circulated sources of traditional media. The languages, covering the humanitarian tragedy, were English, German, Russian and Spanish. The observed time span was a quarter year in summer-autumn 2015. The initial experiment compared the data volume per language of the automatically compiled corpora. The German data volume was considerably higher than those of the other languages, explained by the fact that German speaking countries faced the crisis immediately. The second experiment employed SentiSAIL software tool to perform sentiment analysis per language. The outcome of the experiment was that the dominating sentiment on the English and Spanish corpora was the negative one, whereas on the German and Russian data the neutral one. The final experiment visualized and illustrated the distribution of the positivity and negativity rates among all multilingual sources, revealing a tendency towards clustering per language. In a larger context, these results form part of our contrastive analysis of media coverage of disasters across multiple languages and media.

REFERENCES

[1] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in Proc. of the 1st ACM Conference on Online Social Networks (COSN). Boston, USA: ACM, 2013, pp. 27–38.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques." in Proc. of the ACL conference on Empirical methods in natural language processing (EMNLP), Philadelphia, PA, USA, 2002, pp. 79–86.

[3] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in Proc. of the 5th Conference on Language Resources and Evaluation (LREC06, 2006, pp. 417–422.

[4] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle," in ACL (System Demonstrations), 2012, pp. 115–120.

[5] P. Gonçalves, F. Benevenuto, and M. Cha, "Panas-t: A pychometric scale for measuring sentiments on twitter," CoRR, vol. abs/1308.1857, 2013.

[6] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment Strength Detection in Short Informal Text," J. American Society for Information Science and Technology, vol. 61, no. 12, Dec. 2010, pp. 2544–2558.

[7] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," Journal of Language and Social Psychology, vol. 29, no. 1, 2010, pp. 25–54.

[8] E. Cambria, R. Speer, C. Havasi, and A. Hussain, "SenticNet: A Publicly Available Semantic Resource for Opinion Mining," in AAAI Fall Symposium: Commonsense Knowledge, 2010, pp. 14–18.

[9] P. S. Dodds and C. M. Danforth, "Measuring the happiness of large-scale written expression: songs, blogs, and presidents," Journal of Happiness Studies, vol. 11, no. 4, 2009, pp. 441–456.

[10] G. Backfried et al., "Integration of Media Sources for Situation Analysis in the Different Phases of Disaster Management: The QuOIMA Project," in Proc. of European Intelligence and Security Informatics Conference (EISIC), Uppsala, Sweden, 2013, pp. 143–146.

[11] G. Shalunts and G. Backfried, "SentiSAIL: Sentiment Analysis in English, German and Russian," in Proc. of the 11th International Conference on Machine Learning and Data Mining, ser. MLDM '15, Hamburg, Germany, 2015, pp. 87–97.

[12] G. Backfried et al., "Open Source Intelligence in Disaster Management," in Proc. of the European Intelligence and Security Informatics Conference (EISIC). Odense, Denmark: IEEE Computer Society, 2012, pp. 254–258.

[13] G. Shalunts, G. Backfried, and K. Prinz, "Sentiment analysis of German social media data for natural disasters," in Proc. of the 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM), University Park, Pennsylvania, USA, 2014, pp. 752–756.

[14] R. Remus, U. Quasthoff, and G. Heyer, "Sentiws - a german-language resource for sentiment analysis," in Proc. of the 7th conference on International Language Resources and Evaluation (LREC), Valletta, Malta, 2010, pp. 1168–1171.

[15] S. Momtazi, "Fine-grained german sentiment analysis on social media," in Proc. of the 8th International Conference on Language Resources and Evaluation (LREC). Istanbul, Turkey: European Language Resources Association (ELRA), 2012, pp. 1215–1220.

[16] I. Chetviorkin and N. Loukachevitch, "Extraction of Russian Sentiment Lexicon for Product Meta-Domain," in Proc. of the 24th International Conference on Computational Linguistics (COLING), Bombay, India, 2012, pp. 593–610.

[17] I. Chetviorkin and N. Loukachevitch, "Evaluating Sentiment Analysis Systems in Russian," in Proc. of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria, 2013, pp. 12–17.

[18] A. Moreno-Ortiz, C. Perez-Hernandez, and M. A. Del-Olmo, "Managing multiword expressions in a lexicon-based sentiment analysis system for spanish," in Proc. of the 9th Workshop on Multi-word Expressions, Atlanta, Georgia, USA, 2013, pp. 1–10.

[19] V. P. Rosas, C. Banea, and R. Mihalcea, "Learning Sentiment Lexicons in Spanish," in Proc. of the 8th International Conference on Language Resources and Evaluation (LREC). Istanbul, Turkey: European Language Resources Association (ELRA), 2012, pp. 3077–3081.

[20] J. Brooke, M. Tofiloski, and M. Taboada, "Cross-Linguistic Sentiment Analysis: From English to Spanish," in Proc. of Recent Advances in Natural Language Processing (RANLP). Borovets, Bulgaria: RANLP 2009 Organising Committee / ACL, 2009, pp. 50–54.

[21] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational Linguistics, vol. 37, no. 2, 2011, pp. 267–307.

[22] A. Balahur et al., "Sentiment analysis in the news," in Proc. of the 7th International Conference on Language Resources and Evaluation (LREC). Valletta, Malta: European Language Resources Association (ELRA), 2010.

[23] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis," Computational Linguistics, 2009, pp. 399–433.

[24] G. J. Norman et al., "Current Emotion Research in Psychophysiology: The Neurobiology of Evaluative Bivalence," Emotion Review, vol. 3, no. 3, 2011, pp. 349–359.

 ISBN: 978-1-61208-510-4