Lang Resources & Evaluation (2018) 52:895–919
https://doi.org/10.1007/s10579-018-9413-3

PROJECT NOTES

Annotated news corpora and a lexicon for sentiment analysis in Slovene

Jože Bučar[1,2] · Martin Žnidaršič[4] ·
Janez Povh[2,3]

Published online: 6 February 2018

Abstract In this study, we introduce Slovene web-crawled news corpora with sentiment annotation on three levels of granularity: sentence, paragraph and document levels. We describe the methodology and tools that were required for their construction. The corpora contain more than 250,000 documents with political, business, economic and financial content from five Slovene media resources on the web. More than 10,000 of them were manually annotated as negative, neutral or positive. All corpora are publicly available under a Creative Commons copyright license. We used the annotated documents to construct a Slovene sentiment lexicon, which is the first of its kind for Slovene, and to assess the sentiment classification approaches used. The constructed corpora were also utilised to monitor within-the-document sentiment dynamics, its changes over time and relations with news topics. We show that sentiment is, on average, more explicit at the beginning of documents, and it loses sharpness towards the end of documents.

✉ Jože Bučar
joze.bucar@gov.si

Martin Žnidaršič
martin.znidarsic@ijs.si

Janez Povh
janez.povh@fs.uni-lj.si

[1] Real Estate Mass Valuation System, Surveying and Mapping Authority of the Republic of Slovenia, Ljubljana, Slovenia

[2] Laboratory of Data Technologies, Faculty of Information Studies, Novo mesto, Slovenia

[3] Laboratory for Engineering Design, Faculty of Mechanical Engineering, Ljubljana, Slovenia

[4] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

Keywords News corpus · Sentiment analysis · Lexicon · Annotated corpus · Corpus linguistics · Web-crawling · Word list · AFINN · Slovene · Machine learning · Document classification · Monitoring sentiment dynamics

1 Introduction

The growing interest in the efficient analysis of informal, subjective and opinionated web texts has led to remarkable developments in the field of sentiment analysis. Since 2010, the amount of scientific research on this subject has rapidly and steadily grown. Many studies report on the perception of emotion (i.e. sentiment) in text messages, for example, forecasting the outcomes of elections based on comments found on Twitter and other social media resources, predicting future events and addressing issues of global security.

Data scientists strive to improve the computational understanding of the world's languages. Therefore, the significant increase in the availability and use of language resources for the purpose of computational linguistics in recent years is not surprising. The majority of language resources are in English, but the interest in other languages has increased. Our aspiration is to advance the understanding of sentiment analysis for Slovene and to contribute to the development of its relevant resources.

The main objectives of this study are (1) to describe in detail the acquisition, creation and annotation of various news corpora written in Slovene and (2) to demonstrate how these corpora can be used to improve the sentiment analysis of Slovene texts.

The main contributions of this study are as follows:

- Introduction and detailed description of the procedure used to build annotated web-crawled news corpora, a collection of various news corpora written in Slovene. Three levels of sentiment granularity were used: sentence, paragraph and document levels.
- Development of a lexicon to analyse the sentiment in Slovene texts based on the resources we created. This is the first sentiment lexicon of this type for Slovene.
- Sentiment classification of Slovene texts into two and three classes obtained by two classifiers (Naïve Bayes Multinomial [NBM][1] and Support Vector Machines [SVM]) that performed best in a previous analysis (Bučar et al. 2016). We show in Sect. 5.1 that NBM reached good performance results (F1 score above 90% for a two-class sentiment and above 60% for a three-class sentiment, even in the most demanding evaluation scenarios), which are better than the reported results of related work in comparable experimental settings.
- Analysis of the within-the-document sentiment dynamics of manually annotated documents and the indication that the sentiment is, on average, more pronounced

[1] Multinomial Naïve Bayes is a Bayesian approach that avoids the explicit penalization of nonoccurence of words in documents and is suitable for text classification. Detailed description is out of the scope of this paper, but is available for example in Kibriya et al. (2004) or Aggarwal (2015).

at the beginning of documents and leans to neutrality towards the end of documents.

- Analysis of sentiment dynamics over time and its associations with news topics as an important exemplary use case of the newly developed resources.
- Free access to the newly developed language resources.

The remainder of this paper is organised as follows: Sect. 2 introduces the corpora in Slovene and briefly presents the lexicons for sentiment analysis and annotation. Section 3 gives an overview of corpus construction, followed by the annotation process and ending with an exploration of the corpus. Section 4 describes the lexicon construction and its exploration. Experiments and use cases related to sentiment classification and monitoring of the sentiment dynamics based on the developed language resources are described in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Related work

In this section, we provide a background relating to Slovene corpora. Next, we present lexicons for sentiment analysis, and we describe sentiment analysis and annotation.

2.1 Corpora in Slovene

Slovene is spoken by only two million people, which puts it in the 36th place among the most common languages on the web.[2] The proportion of web pages written in Slovene increased from 0.074 to 0.088% in the past year.

The two largest corpora in Slovene, i.e. slWac (Ljubešić and Erjavec 2011) and Gigafida (Berginc et al. 2013), comprise 1.2 billion words (tokens). The current version of slWac (v2.0) is a web-crawled corpus gathered mostly from the .si domain. It includes tokenization, morphosyntactic descriptions [MSDs][3] and lemmatization with the ToTaLe (Erjavec et al. 2005) tool. Gigafida was obtained from selected texts, written in Slovene, of different genres and styles, mainly from newspapers, magazines and the web.

In addition to these two, many other corpora are available in the CLARIN.SI language resources repository.[4] Examples are KRES (Berginc et al. 2012), a balanced 100-million-word corpus derived from Gigafida, JOS (Erjavec et al. 2010, Erjavec and Krek 2010), a 1-million-word corpus also derived from Gigafida containing lemmas and morphosyntactic descriptions, and JANES (Fišer et al. 2016), a user-generated 167-million-word corpus containing tweets, forums, blogs

[2] https://w3techs.com/technologies/overview/content_language/all.

[3] MSDs are more detailed than is commonly the case for part-of-speech [PoS] tags; they are compact string representations of a simplified kind of feature structures. The first letter of a MSD encodes the PoS. The specifications define the values of the position-determined attributes, for each PoS, its appropriate attributes, their values and one-letter codes.

[4] http://www.clarin.si/.

and comments on news articles and on Wikipedia, ssj500k (Krek et al. 2015), a corpus of 500,000 words manually annotated on the levels of tokenization, sentence segmentation, morphosyntactic tagging, lemmatisation, named entities and, partially, syntactic dependencies, and the IMP corpus of historical Slovene (Erjavec 2014).

CLARIN.SI also contains corpora that are focused on sentiment in text: there is a collection of the IDs of tweets (in 15 languages, including 112,832 IDs of Slovene tweets) with annotated sentiment (Mozetič et al. 2016) and a corpus of 4,777 web commentaries with sentiment annotations (Kadunc and Robnik-Šikonja 2017). Both tweets and commentaries are representatives of very specific and interesting kinds of texts.

2.2 Lexicons for sentiment analysis

Lexicons have been widely used for sentiment and subjectivity analysis, as they represent a simple but effective way to build rule-based opinion classifiers (Perez-Rosas et al. 2012). Lexicon-based approaches are usually domain dependent because the subjectivity of most polarity words is very ambiguous.

One of the first-known, human-annotated lexicons for effect and opinion mining is the General Inquirer lexicon (Stone et al. 1966), which contains 11,788 English words (2291 labelled as negative and 1,915 as positive, with the rest labelled as objective). WordNet is a lexical database for the English language that groups terms into sets of cognitive synonyms (Fellbaum 1998). Several sentiment lexicons are based on WordNet's synonyms, such as SentiWordNet[5], WordNet-Affect[6] and Micro-WNOp.[7] Liu et al. (2005) built a sentiment lexicon based on online customer reviews of products. The sentiment value of the text in English is estimated by adding the sentiment value for each word. Wiebe and Riloff (2005) built OpinionFinder, a large lexicon of clues tagged with prior polarity (negative, neutral and positive), which has grown into MPQA.[8] AFINN (Nielsen 2011) is another commonly used sentiment lexicon containing a list of English words rated with an integer between − 5 (negative) and +5 (positive). A new version ANEW[9], is based on multiple independent labels per item and provides, besides valence, the arousal and dominance for each word.

Sentiment lexicons of various sizes, quality and development methods exist for most Slavic languages; examples are lexicons for Bulgarian (Kapukaranov and Nakov 2015), Croatian (Glavaš et al. 2012), Czech (Veselovská 2013), Macedonian (Jovanoski et al. 2015), Polish (Wawer 2012) and Slovak (Okruhlica 2013).

[5] http://sentiwordnet.isti.cnr.it/.

[6] http://wndomains.fbk.eu/wnaffect.html.

[7] http://www.unipv.it/wnop.

[8] http://mpqa.cs.pitt.edu/.

[9] http://neuro.imm.dtu.dk/wiki/A_new_ANEW:_evaluation_of_a_word_list_for_sentiment_analysis_in_microblogs.

Slovene WordNet (sloWNet) offers a lexical database in Slovene that organises nouns, verbs, adjectives and adverbs in conceptual hierarchies, thereby linking semantically and lexically related concepts (Erjavec and Fišer 2006). Two lexicons for Slovene were derived from the translation of words from English. Martinc (2013) built the AFINN list of words, which is mainly intended for microblogging. Kadunc and Robnik-Šikonja (2016) developed an opinion lexicon with 90,620 negative and positive terms. The research activities on computational linguistics for Slovene correspond to the number of people who speak Slovene. Bearing this in mind, we wish to support the diversity and richness of freely available language resources, as well as the development of Slovene in the future.

2.3 Sentiment analysis

In general, we can divide studies related to sentiment analysis into four main domains: business/financial (e.g. Hatzivassiloglou and McKeown 1997; Das and Chen 2001), film/movie review (e.g. Pang et al. 2002; Taboada et al. 2011), product review (e.g. Turney 2002; Kushal et al. 2003), and political (e.g. Durant and Smith 2006; Ceron et al. 2015). The main sources of such studies are social media (e.g. Colbaugh and Glass 2010; Nakov 2016). Despite the important role that news plays in our lives, the news genre has received much less attention within the sentiment analysis community. However, some studies have explored sentiment analysis to investigate news articles (Balahur et al. 2013). Related to our report, Reis et al. (2015) investigated the sentiment of the business news produced by four major global media corporations and the dynamics of sentiment in news over time.

Although subjectivity in news articles has traditionally tended to be implicit, news stories still have their own biases. The growing trend to foster interactivity and more heavily report the communication of Internet users within the body of news articles is likely to make the expression of subjectivity in news articles even more explicit (Abdul-Mageed and Diab 2011).

Machine-learned sentiment classifiers for tweets in Slovene were developed by Mozetič et al. (2016). Although these models were learned from labelled tweets, the empirical analysis by Fišer et al. (2016) indicated that they might be suitable also for the classification of other types of texts, including news.

2.4 Sentiment annotation

The success of most studies depends on the quality of their knowledge bases—either lexicons containing the sentiment polarity of words or manually annotated corpora for machine learning. Although the majority of manual interventions to assemble lexicons have been performed by bootstrapping (Wiebe and Riloff 2005), it is difficult to bypass the manual annotation process which is often expensive and time-consuming (Hsueh et al. 2009). Sentiment annotation is complex, as it spans the lexical, syntactic and semantic levels. The process of building an annotated corpus is often cyclical, with changes and adjustments to the annotation level and tasks because the data are further examined. Pustejovsky and Stubbs (2012) referred to the annotation process as the MATTER cycle, which involves modelling, annotation, training, testing, evaluation and revision. Several systems, crowdsourcing platforms and tools[10] are used for the retrieval, annotation and analysis of this type of data.

Crowdsourcing is a popular approach to obtain manual annotations (Snow et al. 2008). Hsueh et al. (2009) analysed the data from both expert and non-expert annotators recruited from web services by exploring three selection criteria, i.e. noise level, sentiment ambiguity, and lexical uncertainty, to identify untrustworthy annotators and select suitable items for predictive modelling. They confirmed the utility of these criteria in improving data quality. However, the majority of crowdsourcing problems can be avoided in annotation processes with small and purposely trained groups (Mozetič et al. 2016), such as those in our study (see Sect. 3.2).

3 Constructing annotated news corpora

In this section, we describe the procedure to construct the corpora, a collection of large (more than 3 million words in 10,427 documents) manually annotated news corpora, and the annotation process. We also provide relevant summary statistics about them.

3.1 Corpora construction

The manually sentiment-annotated Slovene news corpus SentiNews 1.0 (Bučar 2017b), which we are introducing in this study, is constructed on the basis of all Slovene news texts with political, business, economic and financial content published between 1st of September 2007 and 31st of December 2013 and retrieved from five widely read Slovene web media resources (24ur[11], Dnevnik[12], Finance[13], Rtvslo[14], Žurnal24[15]). These five web media resources were chosen because they are very popular and have a well-organised digital news archive, facilitating the acquisition of web texts for the selected period.

The text for the corpora was obtained by crawling each of the selected web media resources (Bučar 2017c). Every piece of news was put in a separate text file, containing the official URL of the web medium, the URL of the news, the date of publication of the news, the author, the keywords, the title, the summary and the content of the news. The files are organised in such a way that the hashtag character indicates a new attribute, with each attribute stored in a new line, as shown in Fig. 1.

[10] Apache OpenNLP, GATE, Lydia, MAE, MALLET, MPQA (OpinionFinder 2), Orange, QDA Miner Lite, Phyton (NLTK), R (tm), RapidMiner, TAMS Analyzer, WEKA, etc.

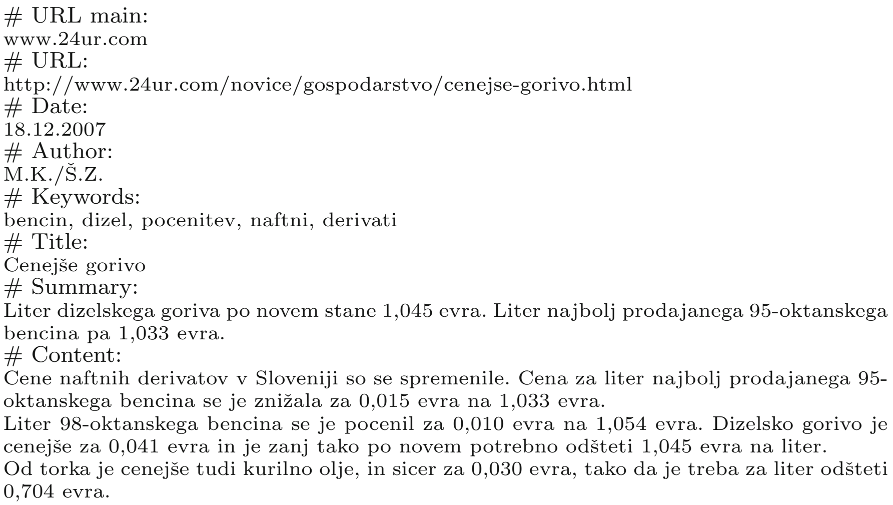
[11] http://www.24ur.com/arhiv/novice/gospodarstvo/.

[12] https://www.dnevnik.si/posel/novice/.

[13] http://www.finance.si/danes/.

[14] http://www.rtvslo.si/gospodarstvo/arhiv/.

[15] http://www.zurnal24.si/archive/slovenija/.



```
# URL main:
www.24ur.com
# URL:
http://www.24ur.com/novice/gospodarstvo/cenejse-gorivo.html
# Date:
18.12.2007
# Author:
M.K./Š.Z.
# Keywords:
bencin, dizel, pocenitev, naftni, derivati
# Title:
Cenejše gorivo
# Summary:
Liter dizelskega goriva po novem stane 1,045 evra. Liter najbolj prodajanega 95-oktanskega
bencina pa 1,033 evra.
# Content:
Cene naftnih derivatov v Sloveniji so se spremenile. Cena za liter najbolj prodajanega 95-
oktanskega bencina se je znižala za 0,015 evra na 1,033 evra.
Liter 98-oktanskega bencina se je pocenil za 0,010 evra na 1,054 evra. Dizelsko gorivo je
cenejše za 0,041 evra in je zanj tako po novem potrebno odšteti 1,045 evra na liter.
Od torka je cenejše tudi kurilno olje, in sicer za 0,030 evra, tako da je treba za liter odšteti
0,704 evra.
```

Fig. 1 Sample format of raw files built by web crawlers written in R

Similarly, within the attribute *Content*, each paragraph is stored separately on a new line. The dates are stored in *dd.mm.yyyy* format.

First, we obtained 217,532 documents published between 1st of September 2007 and 31st of December 2013 (Finance: 110,841, Dnevnik: 47,684, Žurnal24: 39,886, Rtvslo: 10,450 and 24ur: 8671), which were the basis for a sample that we manually annotated (see Sect. 3.2). Second, we subsequently obtained news that was published between 1 January 2014 and 31 January 2016 to estimate the proportion of negative, neutral and positive news (for details see Sect. 5.2). Finally, we stored the data in a MySQL database and developed a web application[16] for the retrieval, storage, annotation and sentiment allocation of Slovene web texts.

3.2 Annotation process

The data retrieval process was followed by cleaning and pre-processing of the data, which included the semi-automatic removal of grammatical and spelling errors. Once we had set up the online environment for the annotation process, six native-speaker annotators[17], were trained in two phases.

[16] http://dejan.amadej.si/test/.

[17] Initially, the project leaders published a call with basic information about the project on the website of the Faculty of Information Studies and its social networks. Several candidates responded to the call. We invited all the candidates to the first meeting, where they were introduced to the objectives of the project, the content and scope of their work. Next, the project leaders selected six candidates, where various criteria were taken into account: (1) candidate's suitability for carrying out the task, (2) candidate's interest, (3) candidate's organisation and (4) gender and age equality. Among all applicants, three women and three men, aged between 19 and 30 and from two different faculties (Faculty of Computer and Information Science in Ljubljana and Faculty of Information Studies in Novo mesto), were chosen to annotate the texts.

In the first phase, they read the basic guidelines for annotation and learned how to use the web application. Together with a referee, they annotated ten news articles on three levels, i.e. document, paragraph and sentence levels, and discussed individual instances. The process of sentiment annotation consists of two sub-processes: comprehension, in which the annotator understands the content, and sentiment judgment, in which the annotator identifies the sentiment. Using a five-level Likert (1932) scale (1—very negative, 2—negative, 3—neutral, 4—positive and 5—very positive), the annotators were told to specify the evoked sentiment with the use of the following instruction: *"Please specify the sentiment from the perspective of an average Slovene web user. How did you feel after reading this news?"*

In the second phase, along with a referee, each of them annotated 50 news items individually. We analysed the agreement among the annotators, which indicated some issues with compound-complex sentences and the influence of the annotators' personal values, beliefs and attitudes. We discussed the instances with lower agreement and resolved the issues that resulted in additional annotation guidelines. In the case of compound-complex sentences with more than one sentiment expressed, such as a journalist's quotes and a politician's statements, we agreed on assigning the one which prevails, or a neutral sentiment in all other cases. We also agreed that the context should be taken into account, but always in accordance with the instructions given.

Finally, the annotators manually annotated a stratified random sample of 10,427 documents independently, i.e. approximately 2000 documents per web medium on the three levels of granularity (Žurnal24: 2212, Rtvslo: 2163, 24ur: 2103, Dnevnik: 2048 and Finance: 2000). Again, they used a five-level Likert scale to annotate documents on the three levels of granularity, and they followed the instructions they were given in the first and second phases. However, each annotator did not manually annotate all the items in the sample. Almost 9% of the news in the sample was annotated by all the annotators, and slightly more than 70% by at least two of them. The sentiment of an instance is defined as the average of the sentiment scores given by the different annotators. An instance was labelled as:

- Negative, if the average of the given scores was less than or equal to 2.4,
- Neutral, if the average of the given scores was between 2.4 and 3.6,
- Positive, if the average of the given scores was greater than or equal to 3.6.

The annotators were paid to provide this annotation service. Annotating the sample took us nearly one year. To evaluate the process of annotation, we explored correlation coefficients by using various measures of inter-annotator agreement at three levels of granularity, as shown in Table 1.

The first four internal consistency estimates of reliability for the scores, shown in Table 1, range between 0 and 1. Values closer to 1 indicate more agreement than values closer to 0. The Cronbach's alpha values indicate very good internal consistency at all levels of granularity. The value of Krippendorff's alpha at the document level of granularity implies a fair reliability test, whereas its values at the paragraph level and sentence level are lower. Fleiss' kappa values illustrate a moderate agreement among the annotators at all levels of granularity. The Kendall's

Table 1 Values of Cronbach's alpha (α_C), Krippendorff's alpha (α_K), Fleiss' kappa (κ) and Kendall's coefficient of concordance (W) between the annotators, as well as the minima (min), maxima (max) and averages (avg) for the Pearson (r_P) and Spearman (r_S) correlation coefficients at the document, paragraph and sentence levels of granularity

	Document level			Paragraph level			Sentence level		
α_C	0.903			0.862			0.856		
α_K	0.691			0.530			0.514		
κ	0.491			0.468			0.454		
W	0.679			0.593			0.586		
	min	max	avg	min	max	avg	min	max	avg
r_P	0.538	0.740	0.628	0.368	0.610	0.514	0.369	0.607	0.501
r_S	0.533	0.744	0.623	0.374	0.609	0.511	0.374	0.612	0.501

values indicate a fair level of agreement between the annotators at all levels of granularity. Correspondingly, the Pearson and Spearman values range from − 1 to 1, where 1 refers to a total positive correlation, 0 to the absence of correlation and − 1 to a total negative correlation. The coefficients show moderate positive agreement among the annotators, but the values decrease when applied to the paragraph and sentence levels. In addition, we observed the correlation between the annotators and found that one of them significantly differs from the rest, which results in overall lower correlations. Our results support the claim by O'Hare et al. (2009) that accurately annotating sentences (or even phrases) can be difficult. In general, the sentiment scores by different annotators are more consistent at the document level than at the paragraph and sentence levels.

The results of our data retrieval are three manually annotated news corpora (SentiNews 1.0) for three levels of granularity with 10,427 annotated documents. The news corpora include different components: document, paragraph or sentence identifier (ID), official URL of the web medium and URL of the news, title, keywords, body (content) of the news, date, reporter's or agency's name (author), manual annotation scores, average and standard deviation of the annotation scores, as well as the sentiment allocation. The identifiers of instances/documents are compatible within these corpora; for example, the document ID in the sentence-based corpus corresponds to the document ID in the document/paragraph-based corpus.

A detailed overview of the corpora is given in Tables 2 and 3. Table 2 presents the relevant statistical information about the manually annotated news corpora. SentiNews 1.0 include 10,427 manually annotated documents at three levels levels of granularity and 214,705 unique words. In Table 3, we report the proportion of instances that we labelled as negative (neg), neutral (neu) and positive (pos) within three levels of granularity for five web media resources. From the annotated news, Finance publishes the most positive news per web medium (22.05%), whereas 24ur publishes the highest proportion of negative news (39.99%). Finance also seems to have a more balanced proportion of negative and positive news compared with the other media.

Table 2 Corpora statistical information

Unit name	Category			Total
	neg	neu	pos	
Documents	3337	5425	1665	10,427
Paragraphs	23,721	51,642	14,636	89,999
Sentences	45,170	96,238	27,491	168,899
Words	1,068,547	1,695,094	497,686	3,261,327
Unique words	107,637	145,889	73,793	214,705
Avg word length (chars)	5.66	5.70	5.71	5.69
Avg sentence length (words)	20.81	20.67	20.25	20.65

Table 3 Proportion (in %) of instances labelled as negative, neutral and positive within each level of granularity in the observed web media

Web medium	Document level			Paragraph level			Sentence level		
	neg	neu	pos	neg	neu	pos	neg	neu	pos
24ur	39.99	43.75	16.26	24.86	64.42	10.72	25.12	64.21	10.67
Dnevnik	32.32	53.08	14.60	25.90	60.11	13.99	25.87	59.88	14.25
Finance	22.60	55.35	22.05	24.74	49.38	25.88	25.85	47.56	26.59
Rtvslo	37.06	49.18	13.76	27.69	58.93	13.38	26.98	60.07	12.96
Žurnal24	27.89	58.59	13.52	27.89	53.53	18.58	29.67	51.04	19.29

Table 4 Attributes, descriptions and data types within JOB 1.0

Attribute	Description	Data type
Word	Headword from the list of Slovene headwords 1.1	String
AFINN freq	Rounded avg_AFINN score	Integer (− 5 to +5)
	Headword frequency (total number of occurrences in the sentence-based news corpus SentiNews 1.0)	Integer (0 to 260,931)
avg_AFINN	Average of AFINN values in the sentence-based news corpus SentiNews 1.0 deducted by the average sentiment of the corpus	Float (− 4.61 to +5.39)
sd_AFINN	Standard deviation of AFINN values in the sentence-based news corpus SentiNews 1.0	Float (0 to 7.071)

4 A lexicon for sentiment analysis in Slovene

In this section, we describe the construction and characteristics of a new lexicon that supports sentiment analysis in Slovene.

Sentiment analysis approaches that use supervised learning commonly depend on sentiment labels. These labels, in most cases, are created manually for a large number of training items, resulting in a costly and time-consuming process.

Alternatively, one can use unsupervised and semi-supervised learning approaches, which commonly rely on the use of lexicons of, for instance, negative and positive terms.

A major advantage of inducing a lexicon directly from data is capturing domain-specific effects. Lexicon-based techniques are also useful in systems for real-time analysis, such as the monitoring of public sentiment towards presidential candidates during election campaigns, for example. Whilst several lexicons for sentiment analysis for English exist, those for Slovene are scarce (see Sect. 2.2), and, to the best of our knowledge, none of them are built directly from a collection of manually annotated texts in Slovene.

4.1 Lexicon construction

The Slovene sentiment lexicon JOB 1.0 (Bučar 2017d), a lexicon for sentiment analysis in Slovene, is constructed on the basis of the list of Slovene headwords 1.1 (Jakopin 2006). JOB 1.0 contains a list of 25,524 headwords from the list, extended with sentiment ratings based on the AFINN model with an integer between -5 (very negative) and +5 (very positive). The ratings are derived from the lemmatised version of the sentence-based news corpus SentiNews 1.0, described in Sect. 3.2.

Table 4, which provides the first insight into the lexicon, contains the attributes, descriptions and data types. The original sentence-level annotations are based on a five-level Likert scale.

The structure of JOB 1.0 is shown in Table 5. For every manually annotated sentence in the corpus, we made a linear transformation of the *avg_sentiment*, i.e. the average of the sentence-based scores (Ann1-Ann6) from the Likert model to the AFINN model to obtain the corresponding AFINN values. Score 1 within the Likert model was transformed to − 5 within AFINN, score 2.4 to − 1.5 (negative sentiment), score 3.6 to 1.5 (positive sentiment) and the score 5 retained.

For every headword in the list, using the annotated sentence-based news corpus, we counted the number of occurrences and calculated the average (*avg_AFINN*) and standard deviation (*sd_AFINN*) of the AFINN values of all the sentences where this headword appeared.

Table 5 Sample format of JOB 1.0. The first line of the lexicon contains the names of the attributes and every headword is stored in a new line along with the associated attributes

Word	AFINN	freq	avg_AFINN	sd_AFINN
a	0	3415	− 0.466	1.787
aa	− 1	57	− 1.116	1.367
ab	− 1	6	− 1.277	2.189
aba	− 1	5	− 0.610	1.046
abančen	0	3	− 0.443	1.443
abc	0	28	0.121	1.564

JOB 1.0 is alphabetically ordered and tab-separated

Because of the different proportions of documents labelled as negative, neutral and positive, the most common words, such as *biti (to be)*, *v (in)*, *in (and)*, had slightly negative *avg_AFINN* values. For this reason, the *avg_AFINN* was deducted by − 0.390, which is the average sentiment of the corpus. Finally, we obtained the AFINN score for every headword in the list by rounding the *avg_AFINN* score.

4.2 Characteristics and illustrative examples

We assigned the *AFINN* score to 25,524 words within the list of Slovene headwords 1.1. The *AFINN* score was not assigned to all the words in the list because some words, such as *aaahah*, *aaajs*, are either not included in the Dictionary of the Slovenian Standard Language[18] or are unusual for our corpora with political, economic and financial content.

Words that contain a negative *AFINN* score indicate a negative sentiment. In a similar way, words indicate a tendency for a positive sentiment if their *AFINN* score is positive. We assigned a negative *AFINN* score to 7898 words and a positive one to 7976 words.

Limiting our observations to two of the most frequent words that express an intensively negative and positive sentiment, *ubiti (to kill)* and *doživetje (experience)*, we used histograms to illustrate the distribution of their frequencies across 11 groups within the AFINN model, as shown in Fig. 2.

Their histograms are both asymmetrical, and the distribution of their frequencies is concentrated either on the negative or the positive side. The word *ubiti (to kill)* has its peak at the *AFINN* score of − 5. By contrast, the word *doživetje (experience)* has its peak at the *AFINN* score of +4. Unsurprisingly, both *ubiti* and *doživetje* arouse a strong sentiment, each in its own way. The word *ubiti* generally provokes a very negative sentiment, and, therefore, we would expect it to be − 5. However, a more detailed inspection reveals its presence in sentences labelled as positive within our corpus, such as *ubiti dve muhi na en mah (to kill two birds with one stone)*. In addition, we present a histogram of the word *delnica (capital stock)*. In contrast to the two previous words, this histogram is symmetrical. It reaches its peak at the *AFINN* score 0 (the frequency is equal to 7024), and provokes neither only a negative nor only a positive sentiment.

5 Experiments and use cases

In this section, we present use cases that showcase the usefulness of the newly developed language resources, particularly the new analytic abilities that are enabled by having sentiment-annotated documents at various levels of granularity. The resources are used to construct and evaluate a collection of sentiment classifiers, estimate the proportions of negative, neutral and positive news within the entire observed web media data and monitor the dynamics of sentiment within documents, over time and in relation to topics in the news.

[18] http://www.fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika/.

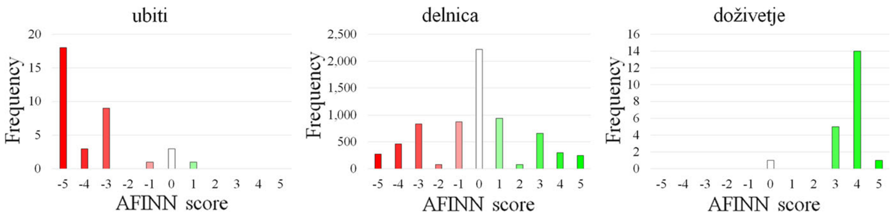


Fig. 2 Frequencies of the words *ubiti* (left), *delnica* (middle), *doživetje* (right) across 11 groups within the AFINN model. The *AFINN* scores are coloured with more intense tones of red and green colour to emphasise sentiment polarity

Access to the resources is provided through the national technology infrastructure for language resources and tools CLARIN[19] and the GitHub[20] website. The resources include information about original texts, metadata and the annotation process. Our resources, available in the CLARIN.SI language resources repository, are not lemmatised, MSD tagged or linguistically processed in any way. However, we lemmatised SentiNews 1.0 with the ToTaLe (Erjavec et al. 2005) tool to derive ratings for the Slovene sentiment lexicon JOB 1.0, described in Sect. 4.1, and to evaluate sentiment-based classification techniques, described in Sect. 5.1 All our resources are available under Creative Commons copyright license Attribution ShareAlike 4.0 International[21] (CC BY-SA 4.0).

5.1 Performance evaluation of sentiment-based classification techniques

Sentiment classification might be the most widely studied problem in the field of sentiment analysis. Most techniques apply supervised learning, in which a bag of words is the most commonly used representation. Here, we empirically evaluate the approaches for two-class (negative and positive) and three-class (negative, neutral and positive) document-based sentiment classification of Slovene news texts.

In our preliminary experiments (Bučar et al. 2016), we studied the performance of various classifiers within the two-class sentiment classification. We have repeated these experiments with the same set of pre-processing options and classifiers with default settings from the WEKA machine-learning toolkit[22] (Witten and Frank

[19] https://www.clarin.si/repository/xmlui/browse?value=Bu%C4%8Dar,%20Jo%C5%BEe&type=author.

[20] https://github.com/19Joey85/Sentiment-annotated-news-corpus-and-sentiment-lexicon-in-Slovene.

[21] https://creativecommons.org/licenses/by-sa/4.0/.

[22] Classifiers and WEKA parameters:

k-Nearest Neighbour (KNN): IBk -K 9 -W 0 -A
“weka.core.neighboursearch.LinearNNSearch -A weka.core.EuclideanDistance -R first-last”
Multinomial Naïve Bayes (NBM): NaiveBayesMultinomial
Support Vector Machine (SVM): SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K
“weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0”
Random Forest (RF): -I 10 -K 0 -S 1
C4.5: J48 -C 0.25 -M 2
Decision Table (DT): -X 1 -S “weka.attributeSelection.BestFirst -D 1 -N 5”
Simple Logistic Regression (SLR): -I 0 -M 500 -H 50 -W 0.0
Voted Perceptron (VP): -I 1 -E 1.0 -S 1 -M 10000

2005), but this time with a balanced data set of 2000 documents (1000 documents labelled as negative and 1000 as positive) from the document-based news corpus SentiNews 1.0. As in the previous study (Bučar et al. 2016), the results show that the NBM and the SVM significantly outperform the other classifiers in terms of classification performance and computational time consumption. Therefore, we focused on these two approaches to assess sentiment classification performance on the newly developed data resources. We tested a large set of pre-processing options, i.e. the term frequency or the term frequency-inverse document frequency weighting scheme, the transformation of upper-case letters to lower case, the removal of stop words, lemmatisation and different combinations of unigrams, bigrams, and trigrams. We performed experiments both with balanced and imbalanced data sets of documents, and classified the documents in two ways to assess the impact of data granularity: (1) based on the average scores of documents and (2) based on the average scores of sentences.

The experiments turned out to be computationally demanding, as it took us one month to build and evaluate the performance of 2400 different predictive models on ten desktop computers simultaneously. We present the best results (considering the pre-processing options) in terms of accuracy and F1-score within the two-class and the three-class document-based sentiment classification for the NBM and the SVM used on the imbalanced data set of documents (see Table 6) and on the balanced data set of documents (see Table 7).

The classifiers perform better on the balanced data, particularly in the three-class scenario. The results indicate that the use of sentence-level granularity is a better option, if available, as this approach, in most cases, yields better results than the document-level one. Overall, the NBM classifier mostly outperforms the SVM, except in terms of accuracy on the imbalanced three-class data. However, accuracy is a less-appropriate measure in imbalanced settings. In terms of pre-processing options, an option shared by all the best solutions is not performing lemmatisation. All but one or two of such options also use transformation to the lower case and stop word removal. The impact of the other options seems to be mixed.

A comparison with a work on a related problem by Mozetič et al. (2016) could only be done in the case of the three-class imbalanced setting for which they report results of about 54% accuracy and 55% F1-score. However, their problem is much more difficult, as they work with tweets (learning and classification), whereas we use news texts, which are longer, more conventional and easier to process.

5.2 Estimating the proportions of negative, neutral and positive news

To estimate the proportions of negative, neutral and positive news up to 2016, we obtained all Slovene news texts that were published between 1st of September 2007 and 31st of January 2016 from the selected web media resources, i.e. 256,567 documents (Finance: 132,986, Dnevnik: 52,417, Žurnal24: 47,735, Rtvslo: 13,420 and 24ur: 10,009).

In this experiment, our goal was to estimate the sentiment in 246,140 remaining documents that were not labelled manually. We applied the NBM predictive model, which was proven as the best within the three-class document-based sentiment

Table 6 Performance evaluation (in %) within the two-class and the three-class document-level sentiment classification for the NBM and the SVM by using five times 10-fold CV with an imbalanced data set of documents (3337 negative, 5425 neutral and 1665 positive)

	Two-class		Three-class	
	NBM	SVM	NBM	SVM
Document-level based on the average scores of documents (based on annotations at the document-level granularity)				
Accuracy	91.07 ± 0.96*	90.68 ± 1.18	64.32 ± 1.21	66.50 ± 1.41*
F1-score	93.19 ± 0.74	93.06 ± 0.88	65.97 ± 1.70*	63.42 ± 1.96
Document-level based on the average scores of sentences (based on annotations at the sentence-level granularity)				
Accuracy	95.21 ± 0.98*	93.10 ± 1.18	66.46 ± 1.49	73.10 ± 1.23*
F1-score	96.38 ± 0.76*	94.86 ± 0.85	61.20 ± 2.21*	55.35 ± 2.31

*Statistically significant (paired *t*-test) at the 0.05 significance level

Table 7 Performance evaluation (in %) within the two-class and the three-class document-level sentiment classification for the NBM and the SVM by using five times 10-fold CV with a balanced data set of documents (1000 negative, 1000 neutral and 1000 positive)

	Two-class		Three-class	
	NBM	SVM	NBM	SVM
Document-level based on the average scores of documents (based on annotations at the document-level granularity)				
Accuracy	92.89 ± 1.65	92.55 ± 1.64	73.09 ± 2.28*	67.94 ± 2.57
F1-score	93.12 ± 1.65*	92.48 ± 1.69	72.77 ± 3.44*	67.71 ± 3.18
Document-level based on the average scores of sentences (based on annotations at the sentence-level granularity)				
Accuracy	97.83 ± 0.98*	96.27 ± 1.34	79.85 ± 1.93*	76.20 ± 2.29
F1-score	97.85 ± 0.97*	96.28 ± 1.34	77.76 ± 3.13*	74.61 ± 3.16

*Statistically significant (paired *t*-test) at the 0.05 significance level

classification in terms of F1-score and time complexity to estimate the proportions of negative, neutral and positive news within the specified web media resources, as shown in Table 8.

Many similarities can be observed in comparison with the results in Table 2. For example, the estimation shows that 24ur and Rtvslo publish the largest proportion of negative news per web medium (24ur: 42%, Rtvslo: 37%), whereas Finance (with 37%) publishes the most positive news. In general, all web media produce much more negative than positive news, with the exception of Finance.

As a result, we obtained another annotated news corpus. The automatically sentiment-annotated Slovene news corpus AutoSentiNews 1.0 (Bučar 2017a) is a large corpus with > 92 million words in 256,567 documents. The structure of the

Table 8 Estimated proportions (in %) of negative, neutral and positive news with political, business, economic and financial content published between 1st of September 2007 and 31st of January 2016 from five Slovene web media resources (n = 256,567)

	Negative	Neutral	Positive	Number of documents
24ur	42.07	37.73	20.20	10,009
Dnevnik	33.37	46.57	20.05	52,417
Finance	19.82	43.23	36.95	132,986
Rtvslo	36.51	49.27	14.23	13,420
Žurnal24	33.90	50.98	15.12	47,735

corpus is very similar to that of the corpora presented in Sect. 3.1. The news corpus includes the following attributes: nid, main URL, URL, title, keywords, content, date, author and sentiment. Unlike in Sect. 3.2, the label (negative, neutral and positive) is estimated using machine-learning techniques.

5.3 Dynamics of sentiment within documents

To determine how sentiment typically changes throughout documents, we investigated the dynamics of sentiment within the sentence-based news corpus SentiNews 1.0. First, we defined a sentiment score for every 10% of the document length on the basis of the linear interpolation between two averaged sentence-based sentiment scores that were closest to our measurement. Second, we averaged all the interpolated sentiment scores for every 10% of the length of a document.

We present the dynamics of the average sentiment and the associated standard deviation through documents, which were labelled as negative, neutral and positive, as shown in Fig. 3. The horizontal axis of the graphs shows the document length from 0% to 100% (for every 10%), whereas the average sentiment and standard deviation, which follow a five-level Likert scale, appear on the vertical axis. The dynamics of the average sentiment of the documents are presented with a coloured line, where the red line refers to the documents manually labelled as negative, the grey to the documents labelled as neutral, and the green to the documents labelled as positive.

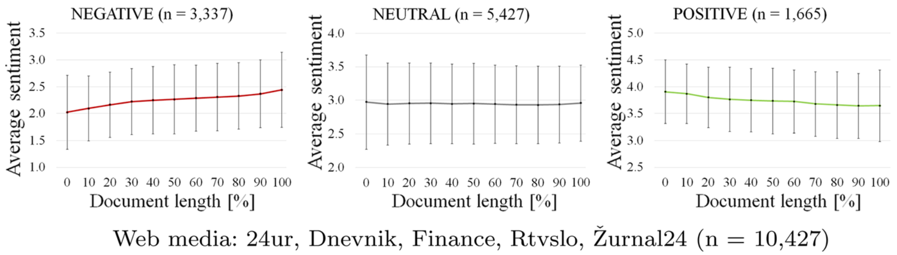


Fig. 3 Dynamics of average sentiment and standard deviation over the length (in %) of documents manually labelled as negative (left), neutral (middle) and positive (right) within the web media

An interesting trend can be observed. The documents labelled as negative, in general, hold a very negative sentiment at the beginning of a document, but steadily lose the intensity of the negative sentiment with the length of the document. A similar trend can be observed in the documents labelled as positive. They also carry a very strong positive sentiment at the beginning of a document, but the sentiment intensity weakens steadily towards the end of the document. However, the average sentiment within the documents labelled as neutral is levelled out. A similar trend within each individual web medium is observed. The observation on the dynamics of sentiment in documents is potentially very important, as it indicates the varied influence of different sections of the document on the overall sentiment. This insight suggests that by using only the starting parts of the news, we might be able to detect sentiment more efficiently and effectively.

5.4 Analysis of sentiment over time

The resources introduced in this paper enable monitoring of sentiment over time to find characteristic patterns and track sudden changes or trends. In an experiment of this kind, we were primarily interested in how the estimated sentiment proportions of negative, neutral and positive news change over time within the individual web medium.

Studies were derived from the automatically sentiment-annotated Slovene news corpus AutoSentiNews 1.0, as described in Sect. 5.2. The results are presented in Figs. 4, 5 and 6. The horizontal axis in the graphs shows different time periods between 1st of September 2007 and 31st of January 2016, whereas the vertical axes present the estimated sentiment proportions (in %) and the total number of documents per time period in the news and within the media.

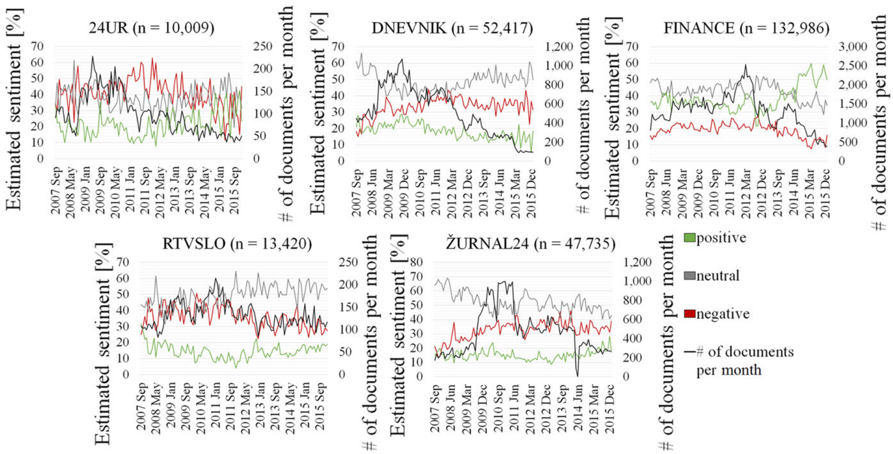


Fig. 4 Estimated sentiment proportion (in %) in the news over time within 24ur (top left), Dnevnik (top middle), Finance (top right), Rtvslo (bottom left) and Žurnal24 (bottom right)

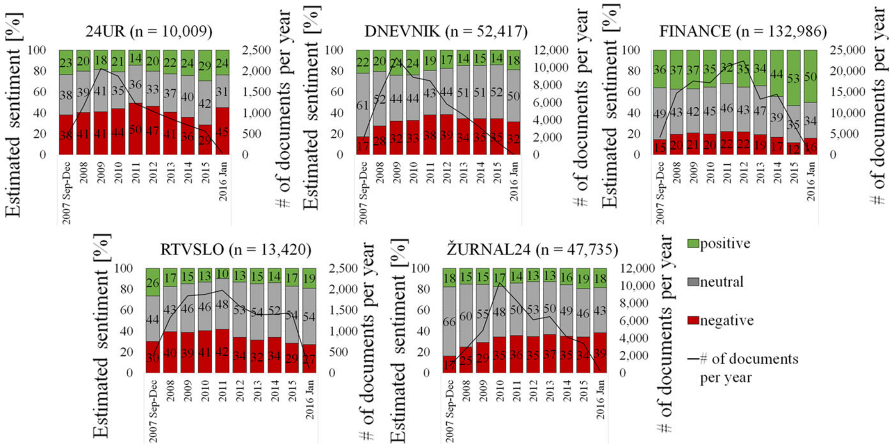


Fig. 5 Estimated sentiment proportion (in %) in the news over years within 24ur (top left), Dnevnik (top middle), Finance (top right), Rtvslo (bottom left) and Žurnal24 (bottom right)

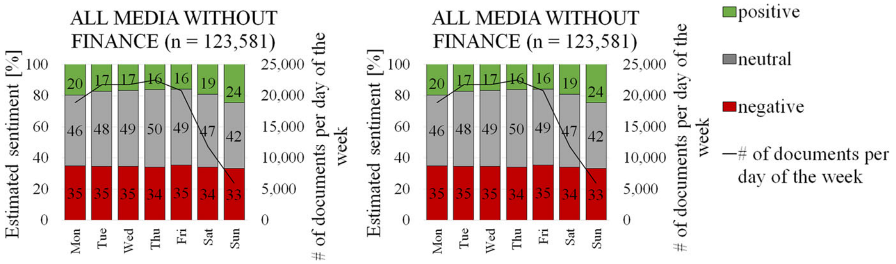


Fig. 6 Estimated sentiment proportion (in %) in the news over days within all media without Finance (left), and in the news published in Finance (right)

The estimated sentiment proportions in the news vary the most within 24ur and the least within Finance. In June 2014, Žurnal24 suffered a dramatic fall in terms of the number of documents (see Fig. 4) as in May 2014, the former owner of Žurnal24 decided to close the company and terminate the website. The new owners of Žurnal24 enabled further publication of news.

In general, most news stories were published in the spring and the least in the summer, especially in August. Thus, the largest amount of news was published by Finance in March (12,488) and the least by 24ur in August (736). 24ur produced the most negative news between 1st of September 2007 and 31st of January 2016, whereas Finance produced the most positive news. Additional observations can be made on the time-associated data. Unsurprisingly, the number of news items published during weekends is much lower than that during other days in the week, but they are obviously more positive (see Fig. 6). The only exception is Finance, which publishes the most negative news during Saturdays. Some negative financial news and events may be deliberately made public late on Fridays after the working hours of stock exchanges.

5.5 Relations of sentiments and topics

Changes in sentiment are usually associated with economic, financial, political, or other events. We analysed the associations of sentiment and topics in the news to further explore the findings from Sect. 5.4 (differences in the news of Finance during weekends) and to verify the usefulness of the new resources for such analyses. For this purpose, we used a popular tool Mallet (McCallum 2002) and the tools and resources that are introduced in this paper.

Our first focus was detection of topics for Finance and all other media separately within weekends (see Fig. 6). For topic retrieval, we defined two parameters: the number of expected topics (3) and the number of words per topic (10). The topics were extracted from the titles of AutoSentiNews 1.0. However, we removed the stop words for Slovene. Our results show that news stories published on Fridays and Sundays within all media without Finance include topics on events, organisations and institutions related to stock exchanges and banks. Moreover, we noticed a positive trend in the news stories published on Sundays within all media without Finance that contain a topic indicating the connection between the economic crisis and hope (topic words: government, Slovenia, companies, help, crisis, etc.). Similarly, news stories that were published on Fridays and Sundays within Finance contain topics related to financial reports, institutions, stock exchanges, banks and cash flows. The major difference is in the news stories published on Saturdays, which include two topics. The first focuses on the foreign economy (topic words: USA, government, Obama, against, growth, forecast, etc.) and the second on the domestic economy (topic words: Pahor, Janša, Janković, Türk, banks, rush, sales, etc.).

The first of the names on the list, which includes topics on the domestic economy, corresponds to the current president of the Republic of Slovenia, Borut Pahor. Analysis of the sentiment and topics related with him was our second major focus. We show the corresponding sentiment dynamics between 1st of September 2007 and 31st of January 2016 in Fig. 7. After the parliamentary election in September 2008, Borut Pahor was appointed as Prime Minister in November 2008.

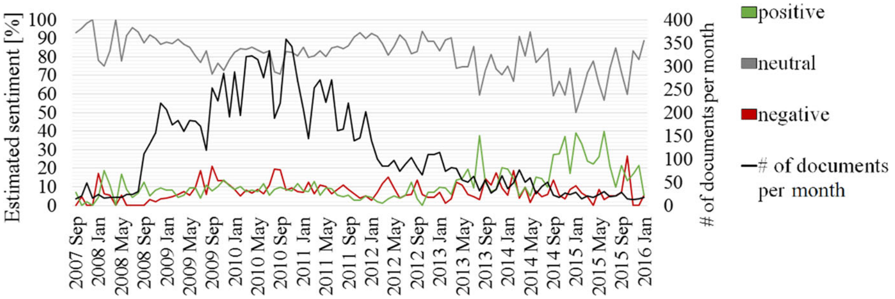


Fig. 7 Dynamics of sentiment in the corpus over time of the current Slovenian president

This phenomenon can be observed in Fig. 7, as his name appears considerably more often than before. In October 2010, Pahor met with the Croatian Prime Minister on the arbitration agreement between the two countries. In September 2011, his government lost a confidence vote in the midst of an economic crisis and of political tensions. Whilst he was the Prime Minister of Slovenia, he was mentioned much more often in the news. Pahor has been the President of Slovenia since December 2012, after a convincing victory at the presidential election. If we focus on the news estimated as negative, we can see that the dynamics of the estimated sentiment proportion are rather weak. The largest proportion of the corresponding negative news was estimated to be published in October 2015. A more detailed insight into topics shows that the news texts mainly focused on issues of the migration crisis and an affair involving a former member of the President's cabinet. However, when dealing with news texts with political, business and financial content estimated as positive, we can observe four peaks. The first can be observed in August 2013 when the President hosted a number of world leaders and business people at the main economic and business conference in Slovenia. The second was in November 2014, which was the result of the strengthening of economic relations with Germany and China. In January 2015, he hosted the President of Qatar and attended several ceremonies and charity events. The last peak can be observed in June 2015 when his commitments to open markets for foreign capital and to an investor-friendly investment climate led to events which connected foreign business people to representatives of leading companies in Slovenia.

We also explored the changes in the sentiments of topics related to the global financial crisis, natural disasters and other highlights in our corpus with an international relevance, and linked them to events (see Fig. 8).

The first chart in Fig. 8, which is related to Lehman Brothers and its bankruptcy, has two peaks. The first, which can be observed in September 2008, is associated with Lehman's bankruptcy filing, followed by the global financial crisis. In September 2009, British television presented the events of the weekend leading to Lehman's bankruptcy in the movie *The Last Days of Lehman Brothers*. In the second chart (see Fig. 8), we explored WikiLeaks, an international organisation that publishes secret and classified information. Here, we can detect three major events. In November 2010, WikiLeaks published a part of leaked diplomatic cables with delicate content labelled as confidential. The second event was detected in September 2011 when an encrypted version of WikiLeaks' US State Department cables was made available via BitTorrent. The last peak in the second chart can be noted in April 2013, when the organisation helped Edward Snowden, who is responsible for the 2013 surveillance disclosures, in leaving Hong Kong. In the third and fourth chart (see Fig. 8), we show that our resources also enable the detection of other events, such as earthquakes and releases of Apple's smartphone iPhone. In the third chart, we explored earthquakes and their occurrence in the corpus. Again, several peaks can be observed. In January 2010, an earthquake struck Haiti with 7.0 Mw; one month later, an earthquake struck Chile with 8.8 Mw. The 2011 earthquake off the Pacific coast of Tohoku was a catastrophic 9.0 Mw undersea earthquake, which triggered powerful tsunami waves and caused great destruction. The last peak in the third chart can be observed in April 2012 when Indonesia was

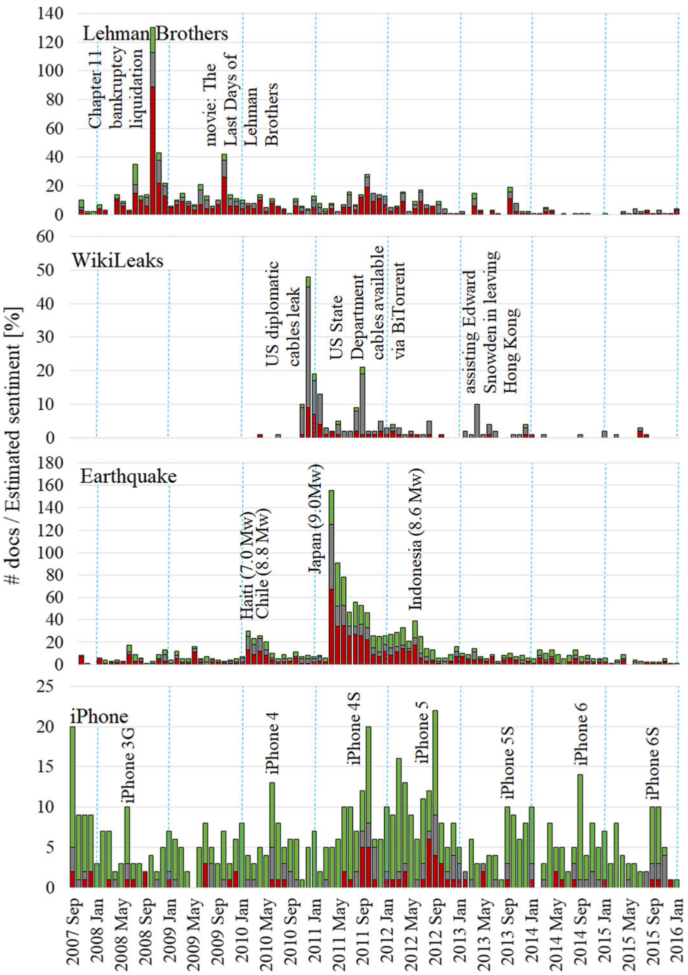


Fig. 8 Dynamics of sentiment in the corpus over time of topics: Lehman Brothers, WikiLeaks, earthquake and iPhone

struck by Indian Ocean earthquakes that reached the highest magnitude of 8.6. The last chart presents the correlation of peaks and releases of new generations of iPhone models. The release of the first-generation model in June 2007 and the announcement of excellent business results in the first quarter of 2012 explain the high values during this period.

6 Conclusions

The rapid growth of information available on the web has increased the interest in the analysis of informal, subjective and opinionated web content. In the past decades, we have witnessed exceptional developments in language technologies, resources and applications.

In this study, we introduce new language resources (corpora, annotations and lexicon) for sentiment analysis in Slovene. We retrieved more than 250,000 news items with political, business, economic and financial content from five Slovene web media resources between 1st of September 2007 and 31st of January 2016. More than 10,000 of them were manually annotated as negative, neutral and positive on three levels of granularity (document, paragraph and sentence levels). Five different measures of correlation were used to evaluate the process of annotation. In general, all the measures indicate good internal consistency at all levels of granularity; however, their values decrease steadily when applied to the paragraph and sentence levels.

In addition, we present some applications and use cases for the obtained resources. First, we empirically evaluated the approaches for the two-class and three-class document-based sentiment classification of Slovene news texts, in which different classifiers and pre-processing options were tested. We have shown that the NBM classifier achieves the best F1-score within the two-class and three-class document-based sentiment classification. Second, we used machine-learning methods to estimate the sentiment in unlabelled documents and to estimate the proportions of negative, neutral and positive news within web media. Third, we provide some illustrative use cases of monitoring the dynamics of sentiment and its relations to the topics expressed in the texts.

An important discovery was found when monitoring the dynamics of sentiment within documents. We observed that sentiment is more emphasised at the beginning of a news story. This discovery might have a significant impact on the practice of sentiment analysis in news.

Using new resources with sentiment classification and topic modelling tools, we made a number of empirical analyses on associations of topics and sentiment in news, which are made possible with the provided corpora.

We hope that other researchers, particularly the representatives of other small language groups, will construct new language resources in a similar way. In the future, we plan to enrich, develop and increase the size of the language resources, coupled with a wider range of media, and, if possible, proceed with a comprehensive manual annotation of news. We intend to further investigate different approaches and techniques that can improve the achieved performance, especially within the three-class, document-based sentiment classification. We believe that lexicon induced directly from data will help capture domain-specific effects, which can improve the achieved performance.

Finally, through our work, we support the open-source community so that future researchers can contribute to the computational linguistics community. For this reason, the developed resources are publicly available.

Acknowledgements This study is based on a work supported by the European Union, The European Regional Development Fund, Slovene Human Resources Development and Scholarship Fund, Ministry of Education, Science and Sport, Slovenia, and the Young Researcher Programme by Slovenian Research Agency. Our research was conducted within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007–2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence. We acknowledge financial support from the Slovenian Research Agency for the research core funding No. P2-0103. We also thank the anonymous reviewers for their comments and suggestions which helped to improve the paper.

References

Abdul-Mageed, M. & Diab, M. T. (2011). Subjectivity and sentiment annotation of modern standard Arabic newswire. In *Proceedings of the 5th linguistic annotation workshop* (pp. 110–118), Portland, OR. Association for Computational Linguistics, Stroudsburg, PA.

Aggarwal, C. C. (2015). *Data mining: The textbook*. New York: Springer.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., et al. (2013). Sentiment analysis in the news. arXiv Preprint ArXiv:1309.6202.

Berginc, N. L., Grčar, M., Brakus, M., Erjavec, T., Holdt, Š. A., Krek, S., et al. (2012). *The Gigafida, KRES, ccGigafida and ccKRES corpora of Slovene language: Compilation, content, use*. Institute for Applied Slovene Studies, Ljubljana: Trojina.

Berginc, N. L., & Ljubešić, N. (2013). Gigafida and slWaC: Topic comparison. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, *1*(1), 78–110.

Bučar, J. (2017a). Automatically sentiment annotated Slovenian news corpus AutoSentiNews 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1109.

Bučar, J. (2017b). Manually sentiment annotated Slovenian news corpus SentiNews 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1110.

Bučar, J. (2017c). R crawlers for five Slovenian web media 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1105.

Bučar, J. (2017d). Slovene sentiment lexicon JOB 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1112.

Bučar, J., Povh, J., & Žnidaršič, M. (2016). Sentiment classification of the Slovenian news texts. In *Proceedings of the 9th international conference on computer recognition systems (CORES 2015)* (pp. 777–787), Wrocław. Springer, Cham.

Ceron, A., Curini, L., & Iacus, S. M. (2015). Using sentiment analysis to monitor electoral campaigns: Method matters-evidence from the United States and Italy. *Social Science Computer Review*, *33*(1), 3–20.

Colbaugh, R. & Glass, K. (2010). Estimating sentiment orientation in social media for intelligence monitoring and analysis. In *Proceedings of the IEEE international conference on intelligence and security informatics (ISI)* (pp. 135–137), Vancouver. IEEE.

Das, S. & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific finance association annual conference (APFA)*, Bangkok.

Durant, K. T. & Smith, M. D. (2006). Mining sentiment classification from political web logs. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, Philadelphia, PA. ACM, New York.

Erjavec, T. (2014). Digital library and corpus of historical Slovene IMP 1.1. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1031.

Erjavec, T. & Fišer, D. (2006). Building Slovene WordNet. In *Proceedings of the 5th international conference on language resources and evaluation (LREC 2006)* (pp. 1678–1683), Genoa. European Language Resources Association.

Erjavec, T., Fišer, D., Krek, S., & Ledinek, N. (2010). The JOS linguistically tagged corpus of Slovene. In *Proceedings of the 7th international conference on language resources and evaluation (LREC 2010)*, Valletta. European Language Resources Association.

Erjavec, T., Ignat, C., Pouliquen, B., & Steinberger, R. (2005). Massive multi lingual corpus compilation: Acquis Communautaire and ToTaLe. *Archives of Control Science*, *15*(4), 253–264.

Erjavec, T. & Krek, S. (2010). Training corpus jos1M 1.1. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1037.

Fellbaum, C., et al. (1998). *WordNet: An electronic database*. Cambridge, MA: MIT Press.

Fišer, D., Smailović, J., Erjavec, T., Mozetič, I., & Grčar, M. (2016). Sentiment annotation of Slovene user-generated content. In *Proceedings of the 2016 conference language technologies and digital humanities (JTDH 2016)* (pp. 65–70), Ljubljana. Faculty of Arts, University of Ljubljana.

Glavaš, G., Šnajder, J., & Bašić, B. D. (2012). Semi-supervised acquisition of Croatian sentiment lexicon. In *Proceedings of the 15th international conference text, speech and dialogue* (pp. 166–173). Springer, Brno.

Hatzivassiloglou, V. & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of ACL and 8th conference of the european chapter of ACL* (pp. 174–181), Madrid. Association for Computational Linguistics, New Brunswick, NJ.

Hsueh, P. Y., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT* (pp. 27–35), Boulder, CO. Association for Computational Linguistics.

Jakopin, P. (2006). List of Slovenian headwords 1.1. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1038.

Jovanoski, D., Pachovski, V., & Preslav, N. (2015). Sentiment analysis in Twitter for Macedonian. In *Proceedings of the international conference on recent advances in natural language processing (RANLP 2015)* (pp. 249–257), Hissar.

Kadunc, K. & Robnik-Šikonja, M. (2016). Analiza mnenj s pomočjo strojnega učenja in slovenskega leksikona sentimenta [Opinion mining using machine learning and Slovene sentiment lexicon]. In *Proceedings of the 2016 conference language technologies and digital humanities (JTDH 2016)* (pp. 83–89), Ljubljana. Faculty of Arts, University of Ljubljana.

Kadunc, K. & Robnik-Šikonja, M. (2017). Opinion corpus of Slovene web commentaries KKS 1.001. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1115.

Kapukaranov, B. & Nakov, P. (2015). Fine-grained sentiment analysis for movie reviews in Bulgarian. In *Proceedings of the international conference on recent advances in natural language processing (RANLP 2015)* (pp. 266–274), Hissar.

Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinomial Naive Bayes for text categorization revisited. In *Australian conference on artificial intelligence* (pp. 488–499), Cairns. Springer.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., & Holz, N. (2015). Training corpus ssj500k 1.4. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1052.

Kushal, D., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th ACM international conference on WWW* (pp. 519–528), Budapest. ACM, New York.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th ACM international conference on WWW* (pp. 342–351), Bremen. ACM, New York.

Ljubešić, N. & Erjavec, T. (2011). HrWaC and slWac: Compiling web corpora for Croatian and Slovene. In *Proceedings of the 14th international conference text, speech and dialogue* (pp. 395–402), Pilsen. Springer.

Martinc, R. (2013). *Measuring sentiment on social network Twitter: Designing a tool and evaluation*. Ljubljana: Faculty of Social Sciences, University of Ljubljana.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PloS One*, *11*(5), 1–26.

Mozetič, I., Grčar, M., & Smailović, J. (2016). Twitter sentiment for 15 European languages. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1054.

Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, A., et al. (2016). Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation*, *50*(1), 35–65.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the 1st workshop on making sense of microposts: Big things come in small packagess* (pp. 93–98), Heraklion.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., et al. (2009). Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the 1st ACM international CIKM workshop on topic-sentiment analysis for mass opinion* (pp. 9–16), Hong Kong. ACM Press, New York.

Okruhlica, A. (2013). Slovak sentiment lexicon induction in absence of labeled data. Master's thesis, Comenius University Bratislava.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of ACL-02 conference on empirical methods in natural language processing* (pp. 79–86), Philadelphia, PA. Association for Computational Linguistics, Stroudsburg, PA.

Perez-Rosas, V., Banea, C., & Mihalcea, R. (2012). Learning sentiment lexicons in Spanish. In *Proceedings of the 8th international conference on language resources and evaluation (LREC 2012)* (pp. 3077–3081), Istanbul. European Language Resources Association.

Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning*. Newton: O'Reilly Media Inc.

Reis, J., Benevenuto, F., de Melo, P. O., Prates, R., Kwak, H., & An, J. (2015). Breaking the news: First impressions matter on online news. arXiv Preprint ArXiv:1503.07921.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254–263), Waikiki, Honolulu. ACM Press, New York.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge: MIT Press.

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

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on ACL* (pp. 417–424), Philadelphia, PA. Association for Computational Linguistics, Stroudsburg, PA.

Veselovská, K. (2013). Czech subjectivity lexicon: A lexical resource for Czech polarity classification. In *Proceedings of the 7th international conference Slovko* (pp. 279–284), Bratislava. RAM-Verlag, Lüdenscheid.

Wawer, A. (2012). Extracting emotive patterns for languages with rich morphology. *International Journal of Computational Linguistics and Applications*, *3*(1), 11–24.

Wiebe, J. & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the international conference on intelligent text processing and computational linguistics* (pp. 486–497).

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Burlington: Morgan Kaufmann.