Towards the Lexicon-Based Sentiment Analysis of Polish Texts: Polarity Lexicon

Konstanty Haniewicz[1], Wojciech Rutkowski[2], Magdalena Adamczyk[3], and Monika Kaczmarek[1]

[1] Department of Information Systems, Faculty of Informatics and Electronic Economy, Poznań University of Economics
{konstanty.haniewicz,monika.kaczmarek}@ue.poznan.pl
[2] Ciber Poland
wojciech.rutkowski@ciber.com
[3] Department of Modern Languages, University of Zielona Góra
m.adamczyk@wh.uz.zgora.pl

Abstract. Due to the increasing amount of information available on the Web, sentiment analysis aiming at an automatic identification of the emotional load of texts is growing in importance. The aim of our research is to devise a reliable method for analsing sentiment in Polish texts, which requires developing adequate polarity lexical resources. In this paper, we discuss a method of building a fine-grained polarity lexicon for Polish based on custom-built review corpora. The compiled lexicon is subsequently tested in the field of sentiment analysis reaching the accuracy level of up to 79%.

Keywords: sentiment analysis, polarity lexicon for Polish, NLP.

1 Introduction and Motivation

Emotions are an important component of texts published in the global Web. The goal of sentiment analysis it to assign to each text "a value expressing an emotional attitude: positive, negative, neutral, objective or bipolar" [22]. Most studies on sentiment analysis use textual material in English as the source of data. However, more and more resources developed specifically for English in order to extract subjective information or identify the polarity of opinions are now successfully being applied to other languages, such as Chinese [25]. Some, though altogether few, attempts of this sort have also been made for the Polish language (see [9]).

Both Polish and English are Indo-European languages; yet, the former is a member of the Balto-Slavic, and the latter of the Germanic language family. The structure of a sentence in Polish is not as constrained as in English and has a freer word order, i.e. parts of speech may be relatively freely rearranged without changing their meaning. As opposed to English, Polish is also a highly inflectional language with robust morphology. Because of these fundamental differences, we deem the task of sentiment classification of the Polish texts more challenging.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 286–295, 2013.

Irrespective of the examined language, there are two main approaches to sentiment analysis, namely the lexicon-based approach and the supervised learning approach. If the former is to be used, the required lexical resources, i.e. dictionaries of sentiment words providing for each item its sentiment score in a given domain, need to be available. While there are numerous polarity resources for the English language (e.g. [3]), those for Polish are scarce and, to the best of our knowledge, none of them is capable of fully supporting sentiment analysis, as shown in the related work section.

The main objective of our work is to build, in an automated manner, a polarity lexicon, i.e. a polarity semantic network, for the Polish language. The quality of the devised system depends strongly on the quality of the developed resources. More specifically, if the designed polarity semantic network does not cover the analysed language model sufficiently well, the results of sentiment analysis will be far from satisfactory. The aim of the study is to advance the understanding of sentiment analysis for the Polish language, thus pointing to its peculiarity, and to contribute to the development of relevant resources for the Polish language that could also be made available to the entire community.

The study follows the pro-active research path based on the design oriented research paradigm [16] and the design science paradigm [6, 7]. The adopted procedure consisted of the following three stages. First, the concept building phase took place, which resulted in establishing a solid theoretical underpining of the work presented in the subsequent sections. The next step was the approach building which involved devising a method of creating a polarity semantic network based on the pre-established theoretical concepts (importantly, the method development drew on a number of the existing approaches to building lexical resources). The last step was concerned with creating the semantic network containing 70000 concepts and running the experiments designed to test the quality of the developed artefact.

The paper is structured as follows. It opens with a brief overview of the relevant research in the domain of interest strictly related to our work. The two subsequent sections are devoted to demonstrating, respectively, the method used for developing the polarity semantic network and the conducted experiments. The paper concludes with final remarks.

2 Related Work

Sentiment analysis refers to "computational treatment of opinion, sentiment and subjectivity in text" [18]. There are two main approaches to sentiment classification, i.e. the supervised learning approach and the lexicon-based approach. Most of the machine learning based approaches use a text classification framework and consist of the following stages: tokenizing, filtering/stemming, training a classifier and, finally, applying the model. The most frequently adopted machine learning approach is the Support Vector Machines with n-gram features trained on a large set of texts with known polarities (usually positive or negative) (e.g. [19] [17]).

In turn, the lexicon-based methods make use of the existing lexical resources. Importantly, the latter are one of the main sources of linguistic information for Natural Language Processing and related areas [20]. The polarity lexical resources differ widely in complexity. They may take the form of simple lists of positive or negative words, as well as that of more complex semantic nets [13]. As word sentiments are domain dependent [23], no general-purpose polarity lexicon is capable of performing reliably for every topic [12].

Creating a polarity lexicon manually is a labour-intensive and error-prone task; in addition, providing a sufficiently widespread coverage of the lexicon is doubtful [12]. Therefore, there are a number of research initiatives aimed at creating polarity lexicons in an automated manner, using either a supervised (e.g. [24]) or unsupervised approach (e.g. [5]). A lot of methods involve utilising other lexical resources, such as thesauri and lexicons, frequently presented in the form of semantic networks. What is more, researchers often use web-documents in order to create polarity lexicons (e.g. [24]). Such lexicons are not limited to any specific word classes and additionally contain slang and multi-word expressions. Experiments have shown that, when compared to lexicons based, for instance, on WordNet [15], they make it possible to achieve a higher performance level in the polarity classification task. Furthermore, there are a few studies aiming at compiling polarity lexicons from corpora for languages other than English which employ automated methods (e.g. [8], [14]). The experiments using those lexicons show about 80% precision with respect to positive and negative polarity on adjectives and adjective phrases.

As far as the English language is concerned, numerous resources are available which successfully cater for the needs of sentiment analysis. Some of them include: SentiWordNet [3] (providing for each synonym a set of three numerical scores, describing respectively how objective (or neutral), positive and negative the terms contained in a synonym set are), WordNet-Affec [21] and ANEW [1]. To the best of our knowledge, there is no polarity lexicon for the Polish language that is readily available to the public; neither have we observed any initiatives to compile one. However, there are some general resources, such as Slowosiec (plWordNet)[1], the biggest Polish wordnet containing 94523 synsets, 133071 lexical units and almost 150000 lexical relations, which may be utilised in the process of developing polarity lexical resources. In addition, a set of tools for language processing (e.g. Morfologik, a morphological analyser) can be used.

In opposition to the few already developed approaches to investigating the emotional overtones of the Polish texts ([2], [9]), the authors of this work adopt the lexicon-based approach and attempt to create the polarity lexical resources for Polish with a view to making them widely available to the public. The accuracy of the constructed polarity semantic network is tested by using it to perform the sentiment classification task.

The mechanism used to create the above mentioned resources as well as the underlying assumptions are presented in the next section of this paper.

[1] `http://nlp.pwr.wroc.pl/en/tools-and-resources/slowosiec`

3 Design of the Mechanism

3.1 Polarity Semantic Network: Main Characteristics

The aim of the study is to provide a dedicated knowledge representation structure suitable for sentiment analysis. After a carefully conducted examination of the available resources, the decision was taken to create a semantic network, whose core is built in an automated manner, based on a number of easily accessible resources, such as dictionaries, thesauri and word lists compiled from the existing open source projects. The main effort while building the semantic network went into formulating a broad set of extraction rules. These rules were applied to the available dictionaries and other resources (such as Wikipedia) in order to identify the relations defined there. Without the structures inherent in dictionaries and other available resources, building a semantic network would be far more difficult. The constructed semantic network, which draws heavily on such structures, was reconfigured and refined thanks to a collective effort of the authors and other volunteers.

The resulting semantic network stores data on over 70000 concepts. While at the moment it is smaller than its alternatives, it has a considerable potential for growth and contains the data unavailable in other known resources. The authors intend to extend the network to over 140000 concepts along with a considerable number of categorised proper names, with a special focus on names connected with brands, organisations and various products.

The structure of the semantic network is designed to store basic data on relationships between individual concepts, such as synonyms, antonyms, hypernyms and homonyms. In addition, every term has a set of additional features, where the most prominent one is the vector storing the data on the sentiment of a concept. In contrast to the existing research, the authors decided to extend sentiment analysis to all available parts of speech. Therefore, if a concept is deemed significant in a given domain, its sentiment analysis score is stored in the semantic network along with the data on the relevant domain. Thus, the polarity semantic network set up in our work may be defined as follows:

The proposed polarity semantic network is a domain-aware sentiment lexicon Ls. Each element in Ls is associated with pairs (s_i, d_j)*, where* s_i *is the sentiment score for a given ith element in a domain* d_j*.*

Although determining the domain of the analysed textual data demands extra effort, the authors strongly believe that it is worthwhile from the language processing point of view.

The main challenge while establishing a domain is to provide a considerable amount of input data that serves as a springboard for a further, fully automated, sentiment identification process. The data that extends the semantic network is gathered from a fully automated analysis of the available corpora. The main tools for the examination of sentiment in textual data include: i) Bayes classifiers, ii) Maximum Entropy, and iii) Support Vector Machines. The result of the application of these tools is a set of terms that can be incorporated directly into the available semantic network. The incorporated terms are those that are most

distinctive with respect to their domain and the actual data (originating from users' reviews).

In addition to the previously discussed aspect, the network contains data on the frequency of any given concept as culled from the reference corpus. This is an important enhancement, as it was observed that many semantic networks suffer from over-specializing when using a network in generalisation or summarising tasks. Providing frequency counts makes it possible for the applied algorithms to promote more popular terms, thus making the generalised or abstracted text more accessible to a human user. While this may not be so important in crude categorisation tasks, it is crucial when evaluating the results of various algorithms by humans.

The constructed semantic network enhanced with the discussed elements is a basic tool used in a number of experiments which aim at offering a solution that is capable of determining the sentiment of a given textual data in an on-line manner. By 'on-line' the authors understand a solution that gives the answer within milliseconds after providing input data.

The following subsections are devoted to a close description of the corpora used in the experiments and additional steps needed to devise the postulated data structure and functionality.

3.2 Data Sets

The data used in sentiment analysis was culled from a number of portals that provide reviews on various goods and services. The total number of the gathered reviews amounted to 356275. Every review, apart from the textual content, was accompanied by a mark provided by its author. The authors of this study made a fundamental assumption about the relation between the provided mark and the content of a given document. It is important to notice that the available scales were different for various sources. It was decided that all the available scales had to be normalised to a scale of 0 (lowest sentiment) to 10 (highest sentiment).

The portals of interest were specialised opinion portals that gathered reviews of particular retail services and products. The majority of opinions offered by the portals are considered to be above average in terms of the score provided by the users. After a careful study of a random sample, it was concluded that a high score is correlated with a high sentiment of the entire review.

The corpus had to be properly balanced to include an equal number of positive and negative reviews. This was done to avoid the predominance of the terms associated with a positive sentiment over the negatively charged ones. A set of sample data is given in Table 1. It has to be emphasised that these are examples of highly positive and highly negative terms.

The data was captured from the following portals: `opineo.pl`, `wizaz.pl`, `opiniuj.pl`, `mgsm.pl`, `cokupic.pl`, `wakacje.pl`. The majority of opinions found there were civil even when a low mark was given by a user. This is a challenge for researchers, as they have to discover the ways of expressing sentiment also in cases were language used to express opinion contains profanity.

Table 1. Sample highly significant positive and negative Polish terms (with approximate English interpretations)

Positive sentiment	Negative sentiment	Neutral sentiment
błyskawiczny [swift]	niesolidny [unreliable]	fajnie [fine]
przystępny [accessible]	nierealny [unreal]	powalać [strike (down)]
znakomity [excellent]	zszargać [bedraggle]	podbić [conquer]
różnorodność [diversity]	przeterminować [expire]	dostateczny [sufficient]
profeska [professional (colloq.)]	ignorancja [ignorance]	spisywać [write down]

3.3 Data Set Preparation

In order to allow for a fair calculation of the sentiment of various concepts, the prepared data set was built with two equally numerous subsets of the original corpus. This was motivated by the fact that a positive sentiment was far more prevalent in the gathered resources (over 95% of the collected reviews were positive). After the preparation of the target sub-corpus, it was divided into two parts. 90% was used to evaluate the sentiment of concepts and the remaining 10% was used as an input for experiments. The total number of the reviews in the balanced corpus amounted to 34265.

As Polish is a highly inflectional language, a procedure was devised that uses Morfologik to prepare a surrogate of an opinion that contains only the basic forms of concepts available in the prepared semantic network. At this stage all stop-words were removed.

The prepared surrogates served as an input for sentiment evaluation. The mean number of traits to be analysed for the balanced sub-corpus equalled 16.8 terms per surrogate.

3.4 Sentiment Calculation

The sentiment for the balanced corpus was calculated with the aid of a specially developed tool that uses Bayes classifiers and SVM in order to define which terms in the corpus are significant for sentiment analysis. The aim of implementing two techniques was to compare the results, yet their effectiveness proved not to diverge from those reported in the literature.

The sentiment value for a given term that was a candidate for inclusion in the semantic network was a number between 0 (lowest sentiment) and 10 (highest sentiment). The output values based on the training set were assigned to the terms in the semantic network to test its applicability.

It is important to notice that an alternative method of sentiment score storage proposed by Liu [10] is also valuable and was used in previous research [4]. This method involves the attachment of a sentiment vector that can store two or three scalars pointing to the inclination of a term towards a positive, negative or neutral sentiment, depending on its context. This is important, as terms are often polysemous and their neighbourhood can define the right scalar to apply in a given context. The total number of the terms selected by the sentiment identification tool amounted to 6685.

4 Experiments

4.1 Experiment Setup

In order to test whether the terms provided by the sentiment identification tool are the appropriate sentiment markers, a test run was carried out. The control sample was composed of 3246 reviews that were not used while establishing potential sentiment markers. A half of the corpus comprised reviews with low scores. The domain of the reviews was the same as the one discussed in the previous section (i.e. products, services, retail reviews).

The reviews were transformed into surrogates with the previously used tools. Out of the 3222 reviews, 2426 surrogates had a sufficient number of terms to calculate a prognosis of the sentiment associated with a given review surrogate. The success rate for those reviews was **78.93%**.

4.2 Discussion of the Results

It is believed that a bigger, balanced corpus should provide more sentiment markers and allow for an even higher success rate. The whole procedure must be repeated for other domains in order to achieve new levels of versatility when applying a semantic network to sentiment analysis. Additional challenges arise from the following traits of human language:

Polysemy and Homonymy. On the level of words alone, and strictly speaking their semantic structure, the highly related phenomena of homonymy and polysemy should not pass unnoticed. Formally, having their component parts identical in sound and spelling, they are indistinguishable. Semantically, while they both rest on double/multiple meanings (in which, in a de-contextualized environment, they may contribute to ambiguity), they mark two distinct types of distance between the meanings they involve.

Idioms/Idiomatic Expressions and Semantic Prosody. On the level of word combinations, the semantic phenomena that should be taken into account include the following: (a) idioms/idiomatic expressions and (b) prosody. In the former case the unique property of semantic structure is non-compositionality, which means that the idiomatic meaning of word strings cannot be derived from the meanings of their component parts (yet, depending on the level of idiomaticity, it can be more or less transparent).

The above mentioned semantic prosody, a relatively new concept introduced in 1993 by Louw [11], is concerned with the fact that the collocates of lexical items are not only random, unrelated words but also semantically defined word classes, which have either a positive or a negative meaning. The prosody of a given word, therefore, depends on the overtones carried by its collocates, and so the verb *powodować* [cause], for instance, can have a negative prosody since it tends to collocate with negatively loaded words, e.g. problem [problem], *wypadek* [accident], *śmierć* [death].

Humour and Irony. Shifting the focus to a deliberate exploitation of both semantic and pragmatic meanings for extra-linguistic purposes, there are at least two phenomena that have to be reckoned with when interpreting the data, namely humour, involving a skilful manipulation of (double/multiple) meanings, and irony, exploiting the interplay of the types of meanings. When it comes to humour, it may have some bearing on the results of the experiment insofar as it depends for its existence on the aforementioned semantic processes of polysemy and homonymy in adjectives and adverbs.

In turn, irony may skew the results as it rests on a calculated mismatch between the literal (explicit/overt) meaning of an utterance, be it a single word or an entire sentence, and the implied (implicit/covert) meaning intended by a speaker.

Grammatical and Lexical Negation. Last but not least, when faced with the data such as those used in the experiment, it needs to be born in mind that there exist structures in language, on the level of grammar and lexis alike, which are capable of producing the exact opposites of the meanings of lexical items. The most straightforward example would be a direct negation in the form of the word *nie* 'not' but less immediately obvious instances thereof were also found in the data gathered for the experiment. These included formulations such as *daleki od (zadowalającego)* [far from (satisfactory)], *mało (ciekawy)* [little (interesting)], *trudno nazwać (pomocnym)* [difficult to call (helpful)] and constructions marking contrast (e.g. *podczas gdy* [whilst], *chociaż/choć* [although], *wprawdzie*, *ale* [but]). The validation proved that a semantic network for Polish extended with domain based sentiment markers is a valuable asset in natural language processing tasks.

5 Conclusions

In this paper we presented a method of building a fine-grained polarity lexicon for Polish on the basis of custom-built review corpora. Moreover, we have tested the obtained lexicon in the field of sentiment analysis. The results of polarity identification are quite promising, with the accuracy rates reaching up to **79%**.

We strongly believe that the created polarity lexicon can be applied without additional automatic or manual processing. However, further enhancements ragarding the number of sentiment markers and supported domains are envisioned.

The future work is intended to be concerned with the development of additional techniques to improve the polarity lexicon in terms of its size and scope. Some effort will also go into introducing an extra dimension nie 'not' in terms of sentiment markers. The authors believe that a homogenised value for every concept in the lexicon reflecting the global sentiment of a concept can be valuable.

Moreover, having accomplished that, the authors would like to further extend the already mastered tools and algorithms in order to offer a coherent solution capable of analysing textual data in terms of local sentiment (at the level of sentence or paragraph, depending on the available context). What is more, the proposed polarity semantic network has a potential for a substantial growth. The growth is secured by the availability of various open access resources (thesauri

and dictionaries). In addition, reorganisation and refinement of the discussed lexicon is an interesting challenge that is also going to be addressed in future research.

References

1. Bradley, M.M., Lang, P.J.: Affective norms for english words (anew): Instruction manual and affective ratings. Psychology, Technical(C-1) (1999)
2. Buczynski, A., Wawer, A.: Shallow Parsing in Sentiment Analysis of Product Reviews. In: Proceedings of the LREC 2008 Workshop on Partial Parsing: Between Chunking and Deep Parsing, pp. 14–18. ELRA, Marrakech (2008)
3. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), pp. 417–422 (2006)
4. Haniewicz, K., Rutkowski, W., Adamczyk, M.: Linguistically Aware Semantic Network for Automated Information Tracking. In: Proceedings of the Eighth International Conference on Signal-Image Technology and Internet-Based Systems, SITIS, pp. 503–509. IEEE Computer Society, Washington, DC (2012)
5. Hassan, A., Radev, D.: Identifying text polarity using random walks. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, pp. 395–403. Association for Computational Linguistics, Stroudsburg (2010)
6. Hevner, A.R.: The three cycle view of design science research. SJIS 19(2), 87–92 (2007)
7. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. Management Information Systems Quarterly 28(1), 75–106 (2004)
8. Kaji, N., Kitsuregawa, M.: Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In: Proceedings of EMNLP-CoNLL, pp. 1075–1083. Prague (2007)
9. Kowalska, K., Cai, D., Wade, S.: Sentiment Analysis of Polish Texts. International Journal of Computer and Communication Engineering 1(1), 39–42 (2012)
10. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (2012)
11. Louw, B.: Irony in the Text or Insincerity in the Writer?: The Diagnostic Value of Semantic Prosodies. In: Baker, M., Francis, G., Toginini-Bognelli, E. (eds.) Text and Technology: In Honour of John Sinclair, pp. 157–176. Benjamins, Amsterdam (1993)
12. Lu, Y., Castellanos, M., Dayal, U., Zhai, C.: Automatic construction of a context-aware sentiment lexicon: an optimization approach. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 347–356. ACM, New York (2011)
13. Maks, I., Vossen, P.: Different Approaches to Automatic Polarity Annotation at Synset Level. In: Proceedings of the First International ESSLLI Workshop on Lexical Resources, pp. 63–71. Publ. online (2011)
14. Maks, I., Vossen, P.: Building a Fine-grained Subjectivity Lexicon from a Web Corpus. In: Chair, N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), pp. 3070–3076. European Language Resources Association (ELRA), Istanbul (2012)

15. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM 38, 39–41 (1995)
16. Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., et al.: Memorandum on design-oriented information systems research. EJIS 20, 7–10 (2011)
17. Paltoglou, G., Thelwall, M.: A study of Information Retrieval weighting schemes for sentiment analysis. Association for Computational Linguistics, pp. 1386–1395 (2010)
18. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. 2(1-2), 1–135 (2008)
19. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, vol. 10, pp. 79–86. Association for Computational Linguistics, Stroudsburg (2002)
20. Sagot, B.: Introduction. In: Proceedings of WoLeR 2011, the 1st International Workshop on Lexical Resources, p. 6. Ljubljana (2011)
21. Strapparava, C., Valitutti, A.: WordNet-Affect: An Affective Extension of WordNet. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, vol. 4, pp. 1083–1086. ELRA, Lisboa (2004)
22. Tromp, E., Pechenizkiy, M.: Senticorr: Multilingual sentiment analysis of personal correspondence. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, ICDMW 2011, pp. 1247–1250. IEEE Computer Society, Washington, DC (2011)
23. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans. Inf. Syst. 21(4), 315–346 (2003)
24. Velikovich, L., Blair-Goldensohn, S., Hannan, K., McDonald, R.: The viability of web-derived polarity lexicons. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 777–785. Association for Computational Linguistics, Los Angeles (2010)
25. Zhang, C., Zeng, D., Li, J., Wang, F.-Y., Zuo, W.: Sentiment analysis of chinese documents: From sentence to document level. J. Am. Soc. Inf. Sci. Technol. 60(12), 2474–2487 (2009)