Evaluation of Most Popular Sentiment Lexicons Coverage on Various Datasets

Justyna Miazga
Pedagogical University of Krakow,
Institute of Computer Science, 2 Podchorazych Ave,
30-084, Krakow, Poland
(+48 12) 662 78 61
justyna.miazga@up.krakow.pl

Tomasz Hachaj
Pedagogical University of Krakow,
Institute of Computer Science, 2 Podchorazych Ave,
30-084, Krakow, Poland
(+48 12) 662 78 66
tomasz.hachaj@up.krakow.pl

ABSTRACT

In this paper we compare three most popular general - purpose lexicon Bing, Afinn and NRC. We omit Loughran on account of its application for financial sentiment analysis. We investigate four various groups of benchmark texts - social media posts, products reviews, songs" lyrics and books. Total count of words is over 2 500 000. This number was than limited to almost 20 thousands after stemming, lemmatization, removing "stop words" and leaving only unique values. As a result, we receive data set composed of literary and universal language which contains lexical, grammatical and stylistic features, accepted and understood by people speaking a specific language. We have gained consistency result between phrases used by Internet users and ones contained in lexicons data between 60 to 75%. The most common part of lexicons and analyzed text has been obtained for books sets, the least for social media post. We also present discussion which lexicon is recommended for various types of dataset taking into account its coverage of used words. Our researchers can be easily reproduced and expanded because we made our source code publicly available.

CCS Concepts

• Computing methodologies→Language resources

Keywords

Sentiment analysis; lexicon-based methods; social media; lyrics; customer review; books.

1. INTRODUCTION

Using a natural language is not complicated, it requires two skills - producing words and understanding it [1]. Recognizing the context and sentiment of speech becomes not trivial task when it is impossible to see paralinguistic features of nonverbal communication. This situation happens when we are analyzing written text. Computer analysis of sentiment is often focused on recognizing positive and negative, sometimes also neutral and other factors. There are many methods that support us in sentiment analysis tasks and which were applied for example in social media platforms [2], songs" lyrics [3], movie reviews [4], airline services [5] or marketing in general [6]. Nowadays the most powerful medium is the Internet, where only small part of sentences has been written grammatically correctly. Because of that it is necessary to face the challenge of understanding speech without observing features of nonverbal communication. Natural Language Processing (NLP) requires constant development. The most important thing is to understand context of sentences. Another thing that needs to be considered understands special signs used by Internet users, like emoticons, overtones that they bring to expressions. Considering difficulties described above, we decided to carry out an analysis of the available research method, which is an analysis sentiment of words based on lexicons. The aim of the research is verification sufficiency of lexicon-based methods for sentiment analysis. For accurate results we have considered different kind of texts, including literary language, song lyrics and informal language.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

SSIP 2019, October 8–10, 2019, Prague, Czech Republic.
© 2019 Association for Computing Machinery
ACM ISBN 978-1-4503-7243-5/19/10…$15.00
https://doi.org/10.1145/3365245.3365251

2. MATERIAL AND METHODS

This study uses lexicon-based method to analyze different groups of texts and compare results to understand which dictionary better coverage tested sentences.

2.1 The Dataset

We have chosen several benchmark datasets that covers different areas of subjects which are in area of interest of sentiment analysis. Those are books, song lyrics, reviews and social media posts. We have used: 6 novels by Jane Austen: "Sense and Sensibility", "Pride and Prejudice", „Mansfield Park", „Emma", „Northanger Abbey" and „Persuasion"; datasets from 5 popular politics and celebrities Twitter accounts; Above 800 song lyrics created by Prince's between 1979 to 2015; Above 23 000 women clothes reviews.

Our choice was not accidental. A selection was dominated by factors: availability of data - all of analyzed datasets are access free, due to this our experiment can be reproduced; various form of expressions covered by our dataset - literary language, lyrics, language used in everyday life and short social media messages; structure of text - arranged (in case of literary and songs) and more chaotic (social media and reviews); count of datasets - large number of words allow to make more accurate results. All of evaluated datasets have been written in English. Twitter is one of the most popular data source used in various researches [7] which allows gathering historical and real time data. We have downloaded data from 5 different Twitter accounts. Accounts have been chosen based on our experiments and information about them. We collected about 13 000 tweets from 5 profiles: Donald J. Trump („Real Donald Trump") - 45th President of the

United States of America; Joanna Krupa - fashion model and celebrity; Oprah Winfrey (Oprah) - TV presenter, producer and actress; Justin Trudeau - 23rd Prime Minister of Canada,; Theresa May - UK Prime Minister and Conservatives Leader. Their common parts include systematic posting, tweets count, the number of followers (people viewing profile) and posts language - English. Despite of fact, that there were more tweets, we are able to collect up to 3,200 tweets per account, which is limited by Twitter. At 24 of November 2018 procedure of the gathering data has been finished.

Table 1. Table presents count of elements that we analyzed in our researchers. All of them are described below.

Type	OD	CI	CAP
Social	230 494	5	4 731
Lyrics	239 789	824	6 691
Books	725 055	6	8 421
Reviews	1 363 325	23 486	869

In table 1 we present numbers of words which we have obtained, where:

- OD is count of words from original dataset,
- CI is count of analyzed items : social - accounts, books - novels, lyrics - songs, reviews - users opinion,
- CAP is number of words after subjecting them to lemmatization, stemming, removing stop words and leaving unique occurrences.

The gathered data contains also 6 novels written by Jane Austen - „Sense and Sensibility”, „Pride and Prejudice”, „Mansfield Park”, „Emma”, „Northanger Abbey” and „Persuasion”. The original data includes 725 055 words. After removing list of „stop words” (definition below) and maintain only unique ones we received data consisting of 8421 items. Songs” lyrics written by Prince have being created since 1975 to 2015 year. We collected data contains 824 lyrics which is consists of 234 458 words, we limited this number to 6691. Original Twitter data consists of 230 494 words and after all irrelevant forms removing is equal to 4731 words. Last analyzed text was women clothes reviews, their count is 23 486 which including 1 363 325 words and we decreased it to 869. Limited number of words is result of removing all of special signs and characters.

The most important common part for used datasets is counts and large variety of words. This attributes allows to collecting interesting various of conclusions. We have sentences which comes from 70”s in lyrics, literary and everyday used language. This multi-items dataset allows making a fair judgment whether lexicon-based methods are sufficient for sentiment analysis.

2.2 Methods

Lexicon based method for sentiment analysis is popular solution often used by scientists and analysts around the world. It brings the expected result during research conducted in the field of understanding human behavior, sales and customers opinion. The lexicon-based approach classifies sentiments from provided data and marks them as positive or negative.

A grammatical correctness of statement requires different forms of words like plural forms of verbs or nouns, such as work, works and working. In fact all of presented words meaning is similar. In this case, obtaining useful results requires evaluating datasets containing only basic form of words, which allows using lexicon based method for sentiment analysis. It means, we have to narrow down an amount of data. To get an expected result we have used methods: stemming, lemmatization, extracting unique words and removing repeating words. Because of that, we have received only not duplicated datasets. Moreover, stemming and lemmatization processes allow acquiring a valid dictionary word which is important to and required by many Information Retrieval systems.

Stemming and lemmatization are part of language modeling techniques. The goal of stemming is to find similar „stem” part to common form. Lemmatization removes inflexional endings and sends back dictionary form of a word. It allows analyzing normalized words form [8].

There are a lot of algorithms to reduce words to their root form. We decided to use package „snowballC” available for R language, which uses the porter stemming algorithm and is compatible with the formula:

$$[C](VC)^{m}[V]$$

Where: C is mark for length one or more consonants, V is mark for length one or more consonants vowels. As a result, you can get 4 forms of words that can easily be described as:

$$[C]VCVC \dots [V] \quad [9]$$

The lemmatization is more complicated process than stemming, it is transformation based on replacement of the grammatical ending by the initial suffix [10]. Lemmatization uses analysis of word morphologic and vocabulary, then tries to removes inflectional endings to return words in a dictionary form [11]. This process is slower than stemming, because of the need to get more knowledge about the context of sentence. In this case we need dictionary and detailed information about parts of speech, including irregular form of them.

Table 2. Example results of lemmatization and stemming.

Original word	lemmatized	stemmed
settled	settle	settl
was	be	wa
respectable	respectable	respect
succession	succession	success

In table 2 we present datasets, where: original word is analyzed word, „lemmatized is word created by lemmatization, stemmed is word created as a result of stemming.

„Stop words” is a list of words which are irrelevant for search engines because they do not change main context of sentences, for example „me”, „our”, „there”, „by”, „between”. Removing them from analyzed text allows to precisely describing of language. It is popular method used and developed in many researchers regarding Natural Language Processing [12].

In this paper we use lexicon-based sentiment analysis with three leading lexicons: „Bing”, „NRC” and “Afinn”. All of them are developed and are base for the analysis of text sentiment. „Bing” was created by Bing Liu and collaborators. It contains 6 788 words, each of them has been binary categorized as positive or negative. NRC by National Research Council Canada consists of 13901 words. In this case, besides of words binary categorization, there are also labels for emotions: fear, trust, sadness, anger, anticipation, joy, disgust, surprise. Afinn lexicon contains 2476 manually rated words with an integer number between -5 (it means that sentiment of sentence is strongly negative) and 5 (which is equal positive sentiment). It was created by Finn Årup Nielsen. We lemmatized and stemmed all of words included in lexicons and left only unique values. As the result we limited count of terms in lexicons to: Bing - 4509, Afinn - 1461, NRC - 5324. After that we compared their content and marked common parts (CP) using Venn Diagram.

2.3 Implementation

The evaluation of dataset presented in Section 2.1 has been made in R v3.5.2. We have used package "dplyr", "tm", and "snowballC" for grammar manipulation. All of diagrams have been drawn using "VennDiagram" package. All lexicons are part of "tidytext". Lemmatization and stemming have been done using "texstem" package. The whole dataset with Jane Austen books is included in the "janeaustenr" package.

3. RESULTS

To visualize common parts of sets from selected data and lexicons (A ∩ B), we have used Venn diagrams. We have calculated coverage of lexicons described in Section 2.2 on various dataset. The information about common parts of words between a lexicon and particular dataset corresponds to the usefulness of individual lexicons in analysis of various types of texts. The more common elements they have, the more descriptive the lexicon becomes. We have prepared diagram which is comparison between content of all lexicons (see Figure 1), which consists of 3 sets with 7 possible intersections. We did it to see how many words are unique for each of sets. This allows for a preliminary judgment, which lexicons have a chance for the highest number of common words with the analyzed text. Fewer mismatches input data and lexicon (words without the indication of sentiment) allows for more accurate results. That would show level of coverage of words contained in dictionaries with obtained collections.

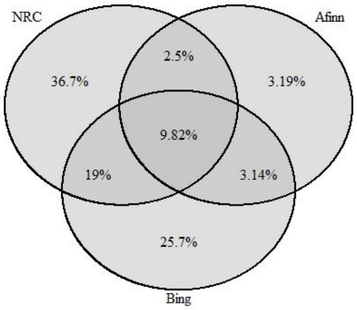

Figure 1. This figure presents common sets of words for all lexicons.

At the figure 1 can be seen:

- intersections of sets marked as "∩", for example NRC ∩ Bing, which means that the resulting collection consists of elements belonging to NRC, Bing and nowhere else;
- union that means all of elements in sets, for example NRC ∪ Bing ∪ Afinn would be all of elements located in selected collections;
- relative complement is marked "\" symbol and contains all elements in one set but not in second one, like NRC \ Afinn is equal to all elements in NRC set without common ones with Afinn.

The union of sets (NRC ∪ Bing ∪ Afinn) is equal to 7828 words but only 769 (9,82%) are in count of elements in their intersection (NRC ∩ Bing ∩ Afinn). About 30% words is the same in Bing and NRC (NRC ∩ Bing). The large number of unique terms is in NRC, above 36%, the lowest in Afinn - 3,19%. Basing on it is expected that NRC will gather most of common words with analyzed texts, next will be Bing and finally Afinn.

To confirm above assumptions, we conducted an attempt to actually evaluate the effectiveness, in this case it will be common part of input dataset with each of lexicons during text analysis. It should be remembered that all words processed in this work have been subjected to lemmatization, stemming and only unique values have left. This applies to analyzed texts as well as lexicons.

Table 3. This table presents percentage results of common parts for analyzed text and individual lexicons, all of lexicons.

Lexicon	Lyrics	Books	Social	Reviews
Bing	17,61%	24,51%	11,96%	21,68%
Afinn	10,58%	11,16%	8,77%	13,58%
NRC	28,78%	33,65%	20,94%	31,53%
CP	3,65%	4,79%	1,61%	00,53%
UCP	66%	58,29%	75%	59,15%

In table 3 are provided information:

- First row contains information about kind of analyzed text (Lyrics, Books, Social, Reviews)
- Bing/Afinn/NRC common part of lexicon and analyzed text resulting from the pattern:
  lexicon ∩ analyzed text
- Common part of text and all lexicons in accordance with the formula:
  CP = NRC ∩ Bing ∩ Afinn ∩ analyzed text
- Uncommon part analyzed text and lexicons calculated as :
  UCP= analyzed text / (NRC ∪ Bing ∪ Afinn)*

*we have removed duplicate elements

Table 3 shows general information about common parts of analyzed text with lexicons. According to our predictions, the NRC, as one of the largest database of words, has the largest number of similar words with those used in analyzed text. It should be remembered that NRC is the lexicon with the highest number of expressions, equal to 13901 in the original version and 5324 after stemming and lemmatization (in case of Bing it is equal to 4509, for Afinn it is 1461). As a result, the table line with the number of lexicon of the common parts and analyzed text for the NRC contains largest values.

From all analyzed texts marked as "books" there is the largest number of common part with all lexicons and the result of uncommon part is the lowest. This effect shows that analysis of sentiments based on the lexicon is very useful in the study of literary texts characterized by the grammatical and stylistic correctness of the language. In the case of "reviews" (data with customers opinion), the number of common parts with lexicons is also high. We can estimate that about one in three words from the analyzed text can be found in the NRC.

Looking at kind of analyzed text we can see that lexicon based method for sentiment analysis is useful source while we are working with properly used language (as it is in literary texts). Correspondingly, the highest number of common parts analyzed text and lexicons is at column "Books", second one are "Lyrics", next "Reviews" and finally "Social". Based on these results, it can be said that the lexicon methods are valuable source of information about the sentiment of statements, mainly in the case of language-correct expressions.

To better understand the data contained in Table 3, we need to take into consideration the detailed information from all intersections of data sets. Therefore in Figures 2-5 we present four Venn diagrams resulting from input data combined with lexicons (Figure 2 – songs" lyrics by Prince, Figure 3 - customer reviews, Figure 4 - tweets, Figure 5 - books by Austen). In this case we have also focused at intersections, relative complement and union of all sets. This time the number of all words is calculated from the formula:

$$\text{NRC} \cup \text{Bing} \cup \text{Afinn} \cup \text{analyzed text},$$

which is equal to 100%. After that we look for intersections between sets. For 4 sets, 15 intersections in the Venn diagram are possible.

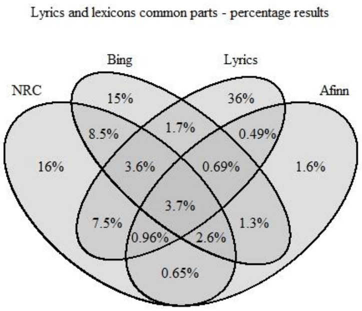

Figure 2. This figure presents common sets of words for all lexicons and text from songs" lyrics made by Prince.

The figure 2 visualize common parts of lexicons Bing, NRC, Afinn and lyrics dataset. The common part of all lexicons and lyrics is 3,7% and it marks elements whose can be found in all of sets. As we predicted before, common part analyzed text with lexicon is highest in case lyrics and NRC - 3,6%, second one is Bing - 1,7% and finally Afinn - 0,49%. This result is calculated in accordance with the formula intersection (∩) of one lexicon and Lyrics with the exclusion (/) of rest. For example number 3,6% is count of all words which are in NRC and lyrics but not in Bing or Afinn. 36% of words used in the prince song lyrics have not been found in any of lexicon. On the other hand, we see that in case of NRC there is 16% of words not used in analyzed text (for Bing it is equal to 15%, for Afinn 1,6%). This number is count of elements which not be found in any other set. Analyzed lyrics by Prince has been created over many years. At this time a lot of factors influencing their content has been changing.

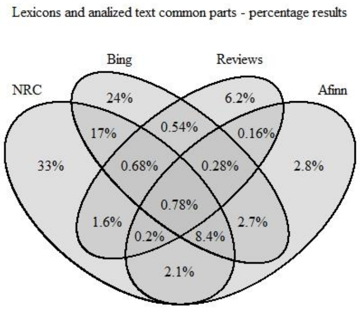

Figure 3. This figure presents common sets of words for all lexicons and text written by customers of clothing store.

The diagram for customer reviews illustrated of figure 3 shows that common part of reviews and all lexicons is 0,78%, uncommon words from analyzed text is only 6,2%. What is really interesting above 50% of lexicons potential is untapped (33% for NRC, 24% for Bing, 2,8% for Afinn). There should be noticed that in this case words count of input data is 896. As a result almost half of words which are available in lexicons did not occur in text. It should be noticed, that elements form review constitutes only 10% of the whole set presented in Figure 3.

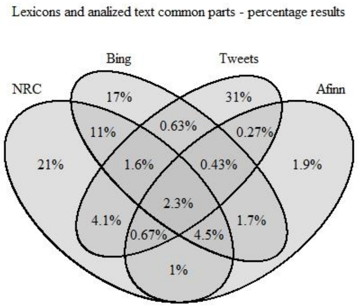

Figure 4. This figure presents common sets of words for all lexicons and posts written by Twitter users.

Figure 4 shows results of Venn diagram for all lexicons and words gathered from Twitter. Their count is 4731, common part for lexicons and tweets is 2,3%. 31% of words from analyzed text have not been found in lexicons. In the case of sentences written by Internet users, it should be remember to prepare the text, which means removing all unmarked characters in lexicons. In this study, we removed all special or irrelevant signs which unchanging contexts of expressions, for example hyperlinks. This was important for our experiment because not removing those elements would give false results.

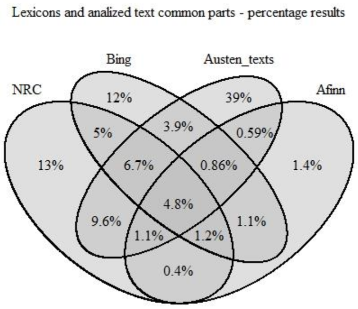

Figure 5. This figure presents venn diagram illustrating common sets of words for all lexicons and text from books written by Austen.

Results illustrated on figure 5 are really interesting because this text has the most common part with all lexicons. What is more, instead of fact that data with books text has the highest number of words (8421) also count of uncommon sentences with lexicons is comparable to results of other analyzed texts. Common part of lexicons and text is 4,8% it is also the highest value in comparing with all achieved results. Unused parts of lexicons is about 33%. Uncommon words number is 39%. Elements which are only in NRC are 13% (Bing - 12%, Afinn - 1,4%). The intersection of Bing and NRC is 5% (counted as intersection of elements from Bing and NRC without Austen_text and Afinn). At this point, it should be recalled that for the remaining groups this result was about 10%.

4. DISSCUSSION AND CONCLUSIONS

Basing of results from 3rd section we can conclude that lexicon-based method for sentiment analysis is important source of knowledge about sentences sentiment. Short text (below 1000 words) achieve worse results than longer texts. Count of words founded in lexicons is higher during analysis of the grammatical and stylistic correctness of the language, what we can observe in case of testing Austen books. Understanding internet users still needs to be improved due to a constant change in the used language. For all four analyzed texts the highest result of common part with lexicon gathered NRC. This is due to number of words which it contained. Because of that we recommend it as the one with the highest coverage of available text. Another advantage of the NRC is categorization of sentiments, which is not only 'positive' or 'negative', but there is also possibility to mark emotions such as anger, expectation, disgust, fear, joy, sadness, surprise, trust. Lexicon-based method is great source of knowledge about sentences sentiment. It is gathering good results for all grammatical and stylistic correct expressions. As a result this method works well in all researches based on literature texts. Results on internet users opinion dataset are satisfying however they could be better, there could be found more common elements of lexicon and analyzed text. Thing which should be improved in this type of lexicon analysis is extensions of lexicons by phrases used especially in social media which contains abbreviation of expressions, for example “gl” would mean good luck. There is still need to understand context of short words like this. The result we obtained are valuable source of knowledge for researchers, marketers, psychologists and marketers. As can be seen, there is a lot of aspects that can be improved. Machine understanding of sentences written in the Internet needs constant improvement. The knowledge about human mood paralinguistic features of nonverbal communication is not easy challenge. Our results proved that lexicon-based method for sentiment analysis is big source of knowledge, mostly for grammatical and stylistic correctness sentences. We decided to analyze the sentiment lexicons coverage on various data. this has not been done before for social media data. Our experiment allowed us to determine the usability indicators of the lexicon-based method, which is the coverage of words in the analyzed set with the lexicon database.

5. REFERENCES

[1] Bateman, J., & Zock, M. 2016. Natural Language Generation. *Oxford Handbooks Online*. DOI= 10.1093/oxfordhb/9780199573691.013

[2] R. Nithish, S. Sabarish, M. N. Kishen, A. M. Abirami and A. Askarunisa, An ontology based sentiment analysis for mobile products using tweets, *2013 Fifth International Conference on Advanced Computing (ICoAC)*, Chennai, 2013, pp. 342-347. DOI= 10.1109/ICoAC.2013.6921974

[3] V. Kumar and S. Minz, "Mood classifiaction of lyrics using SentiWordNet," *2013 International Conference on Computer Communication and Informatics, Coimbatore*, 2013, pp. 1-5. DOI= 10.1109/ICCCI.2013.6466307

[4] M. Wöllmer et al., "YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context," in *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46-53, May-June 2013. DOI= 10.1109/MIS.2013.34

[5] Dey, Nilanjan & Borah, Samarjeet & Ashour, Amira S. & Babo, Rosalina. 2018. Social Network Analytics - Computational Research Methods and Techniques.

[6] Pathak, Xema & Pathak-Shelat, Manisha. 2017. Sentiment analysis of virtual brand communities for effective tribal marketing. *Journal of Research in Interactive Marketing*. 11. 16-38. DOI= 10.1108/JRIM-09-2015-0069.

[7] Nakhasi A ; Bell SG ; Passarella RJ; et al. The potential of Twitter as a data source for patient safety. *J Patient Saf*. 2016 Jan 11; DOI= 10.1097/PTS.0000000000000253

[8] Patel H., Patel B. 2019 Stemmatizer—Stemmer-based Lemmatizer for Gujarati Text. In: Rathore V., Worring M., Mishra D., Joshi A., Maheshwari S. (eds) Emerging Trends in Expert Applications and Security. Advances in Intelligent Systems and Computing, vol 841. *Springer*, Singapore

[9] M.F. 2001, “Snowball: a language for stemming algorithms”, available online: http://snowball.tartarus.org/algorithms/porter/stemmer.html,

[10] R J, Prathibha & M C, Padma. 2015. Design of rule based lemmatizer for Kannada inflectional words. 264-269. DOI= 10.1109/ERECT.2015.7499024.

[11] Balakrishnan, Vimala & Ethel, Lloyd-Yemoh. 2014. Stemming and Lemmatization: A Comparison of Retrieval Performances. Lecture Notes on *Software Engineering*. 2. 262-267. DOI= 10.7763/LNSE.2014.V2.134.

[12] Denny, Matthew & Spirling, Arthur. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*. 26. 1-22. DOI= 10.1017/pan.2017.44002E

[13] Datasets and source code used in this article https://github.com/JusMia/sentiment-analysis