On the Impact of Seed Words on Sentiment Polarity Lexicon Induction

Dame Jovanoski, Veno Pachovski
University American College Skopje
UACS, Macedonia
{jovanoski,pachovski}@uacs.edu.mk

Preslav Nakov
Qatar Computing Research Institute
HBKU, Qatar
pnakov@qf.org.qa

Abstract

Sentiment polarity lexicons are key resources for sentiment analysis, and researchers have invested a lot of efforts in their manual creation. However, there has been a recent shift towards automatically extracted lexicons, which are orders of magnitude larger and perform much better. These lexicons are typically mined using bootstrapping, starting from *very few* seed words whose polarity is given, e.g., 50-60 words, and sometimes even just 5-6. Here we demonstrate that much higher-quality lexicons can be built by starting with hundreds of words and phrases as seeds, especially when they are in-domain. Thus, we combine (*i*) mid-sized high-quality manually crafted lexicons as seeds and (*ii*) bootstrapping, in order to build large-scale lexicons.

1 Introduction

The recent rise of social media has greatly democratized content creation. Facebook, Twitter, Skype, WhatsApp and LiveJournal are now commonly used to share thoughts and opinions about anything in the surrounding world. This proliferation of social media content has created new opportunities to study public opinion, with Twitter being especially popular for research due to its scale, representativeness, variety of topics discussed, as well as ease of public access to its messages.

Naturally, this abundance of data has attracted business and research interest from various fields including marketing, political science, and social studies, among many others, which are interested in questions like these: *Do people like the new Apple Watch? What do they hate about iPhone6? Do Americans support ObamaCare? What do Europeans think of Pope's visit to Palestine? How do we recognize the emergence of health problems such as depression? Do Germans like how Angela Merkel is handling the refugee crisis in Europe? What do republican voters in USA like/hate about Donald Trump?* Answering these questions requires studying the sentiment of opinions people express in social media, which has given rise to the fast growth of the field of sentiment analysis in social media.

Initially, sentiment analysis was addressed as a text classification problem, but it was soon realized that sizable performance gains can be obtained from using carefully built sentiment polarity lexicons as a source of external knowledge. Thus, researchers have invested a lot of efforts in the manual creation of such lexicons, which were typically of small to moderate size, e.g., less than 10,000 words. Recently, there has been a shift towards using automatically extracted lexicons, which are orders of magnitude larger and perform much better. These lexicons are typically mined using bootstrapping, starting from *very few* seed words whose polarity is given, e.g., 50-60 words, and sometimes even just 5-6.

Here, we demonstrate that sizable further performance gains can be observed by starting with mid-sized seeds (hundreds of words and phrases), thus getting the best of both worlds: (*i*) using high-quality mid-sized manually crafted lexicons as seeds, and (*ii*) extending them automatically using bootstrapping.

The remainder of the paper is organized as follows: Section 2 presents some related work. Section 3 describes our training and testing datasets. Section 4 presents the various lexicons we created for Macedonian. Section 5 gives details about our system, including the pre-processing steps and the features used. Section 6 describes our experiments and discusses the results. Section 7 concludes with possible directions for future work.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,
pages 1557–1567, Osaka, Japan, December 11-17 2016.

2 Related Work

In this section, we first present work on sentiment analysis in general: methods used, work on sentiment analysis on Twitter, and relevant tasks at SemEval. Then, we present work on sentiment polarity lexicon induction, and finally, we discuss sentiment analysis for Macedonian.

2.1 Sentiment Analysis

Research in sentiment analysis started in the early 2000s. Initially, it was regarded as standard document classification into topics (Pang et al., 2002). However, researchers soon realized that it was quite different from standard document classification (Sebastiani, 2002), e.g., into categories such as business, sport and politics, and that sentiment analysis crucially needs external knowledge in the form of sentiment polarity lexicons. See for example the surveys by Pang and Lee (2008) and Liu and Zhang (2012) for more detail about research in sentiment analysis.

Around the same time, other researchers realized the importance of external sentiment lexicons, e.g., Turney (2002) proposed an unsupervised approach to learn the sentiment orientation of words/phrases: positive vs. negative. Later work studied the linguistic aspects of expressing opinions, evaluations, and speculations (Wiebe et al., 2004), the role of context in determining the sentiment orientation (Wilson et al., 2005), of deeper linguistic processing such as negation handling (Pang and Lee, 2008), of finer-grained sentiment distinctions (Pang and Lee, 2005), of positional information (Raychev and Nakov, 2009), etc. Moreover, it was recognized that in many cases, it is crucial to know not just the polarity of the sentiment, but also the topic towards which this sentiment is expressed (Stoyanov and Cardie, 2008).

Early sentiment analysis research focused on customer reviews of movies, and later of hotels, phones, laptops, etc. Later, with the emergence of social media, sentiment analysis in Twitter became a hot research topic. Unfortunately, research in that direction was hindered by the unavailability of suitable datasets and lexicons for system training, development and testing. While some Twitter-specific resources were developed, initially they were either small and proprietary, such as the i-sieve corpus (Kouloumpis et al., 2011), were created only for Spanish like the TASS corpus (Villena-Román et al., 2013), or relied on noisy labels obtained automatically based on emoticons and hashtags (Mohammad, 2012; Pang et al., 2002; Mohammad et al., 2013).

This situation changed with the shared task on *Sentiment Analysis on Twitter*, which was organized at SemEval, the International Workshop on Semantic Evaluation, a semantic evaluation forum previously known as SensEval. The task ran in 2013, 2014, 2015 and 2016, attracting over 40+ of participating teams in all four editions. While the focus was on general tweets, the task also featured out-of-domain testing on SMS messages, LiveJournal messages, as well as on sarcastic tweets.

SemEval-2013 task 2 (Nakov et al., 2013) and SemEval-2014 Task 9 (Rosenthal et al., 2014) focused on expression-level and message-level polarity. SemEval-2015 Task 10 (Rosenthal et al., 2015; Nakov et al., 2016b) featured topic-based message polarity classification, on detecting trends towards a topic, and on determining the out-of-context (a priori) strength of association of Twitter terms with positive sentiment. SemEval-2016 Task 4 (Nakov et al., 2016a) introduced a 5-point scale, which is popular and is commonly used for human review ratings on popular websites such as Amazon, TripAdvisor, Yelp, etc.; from a research perspective, this meant moving from classification to *ordinal regression*. Moreover, some subtasks of the general task focused on *quantification*, i.e., determining what proportion of a set of tweets on a given topic are positive/negative about it. It also featured a 5-point scale *ordinal quantification* subtask (Gao and Sebastiani, 2015).

Other related (mostly non-Twitter) tasks explored aspect-based sentiment analysis (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016), sentiment analysis of figurative language on Twitter (Ghosh et al., 2015), implicit event polarity (Russo et al., 2015), stance in tweets (Mohammad et al., 2016), out-of-context sentiment intensity of phrases (Kiritchenko et al., 2016), and emotion detection (Strapparava and Mihalcea, 2007). Some of these tasks featured languages other than English.

2.2 Sentiment Polarity Lexicons

Despite the huge variety of knowledge sources explored in the literature, sentiment polarity lexicons remained the only universally recognized resource for the task of sentiment analysis. Until recently, such sentiment polarity lexicons were manually crafted, and were thus of small to moderate size, e.g., LIWC (Pennebaker et al., 2001) has 2,300 words, the General Inquirer (Stone et al., 1966) contains 4,206 words, Bing Liu's lexicon (Hu and Liu, 2004) includes 6,786 words, and MPQA (Wilson et al., 2005) has about 8000. Early efforts in building them automatically also yielded lexicons of moderate sizes such as the SentiWordNet (Esuli and Sebastiani, 2006; Baccianella et al., 2010).

However, recent results have shown that automatically extracted large-scale lexicons (e.g., up to a million words and phrases) offer important performance advantages, as confirmed at shared tasks on Sentiment Analysis on Twitter at SemEval 2013-2016 (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov et al., 2016a), where over 40 teams participated four years in a row.

Using such large-scale lexicons was crucial for the performance of the top-performing systems. Similar observations were made in the related Aspect-Based Sentiment Analysis task at SemEval 2014 (Pontiki et al., 2014). In both tasks, the winning systems benefitted from building and using massive sentiment polarity lexicons (Mohammad et al., 2013; Zhu et al., 2014).[1] The two most popular large-scale lexicons were the Hashtag Sentiment Lexicon and the Sentiment140 lexicon, which were developed by the team of NRC Canada for their participation in the SemEval-2013 shared task on sentiment analysis on Twitter.

The importance of building sentiment polarity lexicons has resulted in a special subtask (Rosenthal et al., 2015) at SemEval-2015 (part of Task 4), and an entire task (Kiritchenko et al., 2016) at SemEval-2016 (namely, Task 7), on predicting the out-of-context sentiment intensity of words and phrases.[2]

These large-scale automatic lexicons are typically built using bootstrapping, starting with a small set of seeds of, e.g., 50-60 words, and sometimes even just two emoticons (Mohammad et al., 2013).

Here, we demonstrate that sizable further performance gains can be observed by starting with mid-sized seeds (i.e., hundreds of words and phrases), thus getting the best of both worlds: (*i*) using high-quality mid-sized manually crafted lexicons as seeds, and (*ii*) further extending them automatically using bootstrapping.

2.3 Sentiment Analysis for Macedonian

In our experiments below, we focus on Macedonian (tweets), for which we only know two publications on sentiment analysis, none of which is about Twitter.

Gajduk and Kocarev (2014) experimented with 800 posts from the Kajgana forum (260 positive, 260 negative, and 280 objective), using SVM and Naïve Bayes classifiers, and features such as bag of words, rules for negation, and stemming.

Uzunova and Kulakov (2015) experimented with 400 movie reviews[3] (200 positive, and 200 negative; no objective/neutral), and a Naïve Bayes classifier, using a small manually annotated sentiment lexicon of unknown size, and various preprocessing techniques such as negation handling and spelling/character translation.

Unfortunately, the datasets and the generated lexicons used in the above work are not publicly available, and/or are also from a different domain (i.e., not Twitter). As we are interested in sentiment analysis of Macedonian tweets, we had to build our own datasets. We have described these datasets and initial experiments with them in an earlier publication (Jovanoski et al., 2015), where the focus was on the datasets and on the classifier; in contrast, here we focus on assessing the impact of our proposed lexicon generation method. Below we will describe these datasets in detail, for the sake of self-containment of the present paper.

[1] Such lexicons proved useful for other tasks at SemEval, e.g., for SemEval-2016 Task 3 on Community Question Answering (Balchev et al., 2016).

[2] We should note though that the utility of using sentiment polarity lexicons for sentiment analysis probably needs to be revisited, as the best system at SemEval-2016 Task 4 could win without using any lexicons (Deriu et al., 2016).

[3] There have been also experiments on movie reviews for the closely related Bulgarian language (Kapukaranov and Nakov, 2015), but there the objective was to predict user rating, which was addressed as an ordinal regression problem.

3 Data

We downloaded half a million tweets in Macedonian, which we collected over a six-month period spanning from November 2014 to April 2015. We tried to download all Macedonian tweets based on the Twitter language classification. However, it turned out that in many cases, the returned tweets were in Bulgarian or Russian, which are also Slavic languages and share the same alphabet with Macedonian. Thus, we trained and used our own Naïve Bayes classifier, which achieved over 95% accuracy.[4] We used part of these tweets as training and testing data, and the rest for building automatic lexicons.

Table 1 shows statistics about the training and the testing datasets. We can see that they are somewhat balanced between positive and negative tweets, and that there is smaller proportion of neutral tweets.[5]

The *testing data* was annotated for sentiment at the tweet level (using *positive*, *negative*, and *neutral/objective* as labels[6]) by two annotators, both native speakers of Macedonian. The Cohen's Kappa statistics (Cohen, 1960) for the inter-annotator agreement was 0.64, which corresponds to *substantial* agreement (Landis and Koch, 1977). Our follow-up analysis has shown that the main disagreement was about distinguishing between negative and neutral tweets. In the final testing dataset, we discarded all tweets with disagreement (a total of 482 tweets).

The *training data* was annotated by a single annotator, one of those who annotated the testing dataset. In addition to producing tweet-level sentiment polarity annotations, the annotator further marked the positive and the negative phrases inside each tweet. We will use the set of these words and phrases, together with their polarities, as a sentiment lexicon, and also as seeds when bootstrapping a large-scale automatic sentiment lexicon from the remaining unannotated tweet messages.

Dataset	Positive	Neutral	Negative	Total
Train	2,610 (30%)	1,280 (15%)	4,693 (55%)	8,583
Test	431 (38%)	200 (18%)	508 (44%)	1,139
No annot.	–	–	–	0.5M

Table 1: Statistics about the datasets.

4 Sentiment Lexicons

A sentiment lexicon contains words and phrases annotated with positive and negative sentiment, sometimes with numerical intensity, e.g., *spectacular* could have positive strength of 0.91, while for *okay* that might be 0.3. Below we describe the sentiment lexicons we built and experimented with.

4.1 Manually-Crafted Lexicon

As we mentioned above, our training dataset was annotated with sentiment words and phrases, a total of 1,088: 459 positive and 629 negative. These terms form our manually-crafted lexicon.

4.2 Translated Lexicons

As no sentiment polarity lexicons are publicly available for Macedonian, we translated some popular English manually-crafted lexicons such as Bing Liu's lexicon (2,006 positive and 4,783 negative), and MPQA (2,718 positive and 4,912 negative), and a Bulgarian lexicon (5,016 positive and 2,415 negative), extracted from a movie reviews website (Kapukaranov and Nakov, 2015), which includes 694 positive and 2,966 negative English words. We used Google Translate, and we further manually corrected some of the results: we removed some bad translations and we corrected the grammar.

[4] At the 2015 Discriminating between Similar Languages (DSL) shared task (Zampieri et al., 2015), the participating systems distingushed Macedonian from Bulgarian with 100% accuracy, which shows that this is an easy task; as a result, this language pair was not included in the 2016 edition of the task (Malmasi et al., 2016). Our Naïve Bayes classifier achieved slightly lower accuracy as we deal with tweets, which are short and harder to categorize than the newswire texts in the DSL task.

[5] It was previously reported that most tweets are neutral, but this was for English, and for tweets about selected topics (Rosenthal et al., 2014). Here, we have no topic restriction. Moreover, there is an ongoing political crisis in Macedonia, and thus Macedonian tweeps express a lot of emotions rather than staying neutral.

[6] Following (Nakov et al., 2013), we merged *neutral* and *objective* as they are commonly confused by annotators.

4.3 Bootstrapped Lexicons

Various approaches have been proposed in the literature for bootstrapping sentiment polarity lexicons starting from a small set of seeds: positive and negative terms (words and phrases).

A very influential approach is that of Turney (2002), which uses pointwise mutual information and bootstrapping to build a large lexicon and to estimate the semantic orientation of each word in that lexicon. The idea is to start with a small set of seed positive (e.g., *excellent*) and negative words (*bad*), and then to use these words to induce sentiment polarity orientation for new words in a large unannotated set of texts (in his case, product reviews). The idea is that words that co-occur in the same text with positive seed words are likely to be positive, while those that tend to co-occur with negative words are likely to be negative. To quantify this intuition, Turney defines the notion of sentiment orientation (SO) for a term w as follows:

$$SO(w) = pmi(w, pos) - pmi(w, neg)$$

where PMI is the pointwise mutual information, pos and neg are placeholders standing for any of the seed positive and negative terms, respectively, and w is a target word/phrase from the large unannotated set of texts (here tweets).

A positive/negative value for $SO(w)$ indicates positive/negative polarity for w, and its magnitude shows the corresponding sentiment strength. In turn, $pmi(w, pos) = \frac{P(w,pos)}{P(w)P(pos)}$, where $P(w, pos)$ is the probability to see w with any of the seed positive words in the same tweet,[7] $P(w)$ is the probability to see w in any tweet, and $P(pos)$ is the probability to see any of the seed positive words in a tweet; $pmi(w, neg)$ is defined similarly.

The pointwise mutual information (PMI) is a notion from information theory: given two random variables A and B, the mutual information of A and B is the "amount of information" (in units such as bits) obtained about the random variable A, through the random variable B (Church and Hanks, 1990).

Let a and b be two values from the sample space of A and B, respectively. The *pointwise* mutual information between a and b is defined as follows:

$$pmi(a; b) = \log \frac{P(A = a, B = b)}{P(A = a) \cdot P(B = b)} = \log \frac{P(A = a|B = b)}{P(A = a)} \tag{1}$$

$pmi(a; b)$ takes values between $-\infty$, which happens when $P(A = a, B = b) = 0$, and $\min\{-\log P(A = a), -\log P(B = b)\}$, when $P(A = a|B = b) = P(B = b|A = a) = 1$.

In his experiments, Turney (2002) used five positive and five negative words as seeds. His PMI-based approach further served as the basis for the creation of the two above-mentioned large-scale automatic lexicons for sentiment analysis in Twitter for English, initially developed by NRC for their participation in SemEval-2013 (Mohammad et al., 2013). The *Hashtag Sentiment Lexicon* uses as seeds hashtags containing 32 positive and 36 negative words, e.g., `#happy` and `#sad`. Similarly, the *Sentiment140* lexicon uses smileys as seed indicators for positive and negative sentiment, e.g., `:)`, `:-)` and `:))` as positive seeds, and `:(` and `:-(` as negative ones.

Recently, Severyn and Moschitti (2015) proposed an approach to lexicon induction, which, instead of using PMI **(SO)**, assigns positive/negative labels to the unlabeled tweets (based on the seeds), and then trains an SVM classifier on them, using word n-grams as features. These n-grams are then used as lexicon entries with the learned classifier weights as polarity scores.

In our experiments below, we calculate $SO(w)$ using PMI or LR (Logistic Regression[8]), and we experiment with different seeds:

- the 1+1 seeds of Turney (2002), translated to Macedonian ("excellent" and "poor");
- the 7+7 seeds of Turney and Littman (2003), translated to Macedonian;

[7] Here we explain the method using tweets as this is how we are using it, but Turney (2002) actually used page hits in the AltaVista search engine.

[8] LR worked better than SVM in our experiments.

- our 5+5 seeds, manually selected words in Macedonian with strong sentiment;
- 30+30 seeds, which we obtained by translating the 32+36 seeds[9] used for the *Hashtag Sentiment Lexicon* lexicon (but we used these seeds as regular words, not as hashtags);
- the 3+2 smileys from above;
- the 459+629 terms from our manually-crafted lexicon (we further experiment with random proportional positive/negative subsets of 100, 200, and 500 words thereof).

Table 2 shows some statistics about the lexicons we built (unigrams + bigrams) on the unanotated 0.5M tweets. We can see that the larger the seed, the larger the bootstrapped lexicons. Note that the lexicon sizes for PMI and LR are the same as they are calculating the sentiment orientation for the exactly same terms; what differs is the way the weights are being calculated.

Type of seed	**Seeds**	**Unigrams**	**Bigrams**	**Total**
Smileys: NRC	5	128	2,163	2,291
Words: Turney	10	865	14,343	15,208
Words: NRC	60	1,669	32,459	34128
Words: MCL	100	1,926	40,242	42,168
Words: MCL	200	3,752	60711	64,463
Words: MCL	500	7,219	124,977	132,196
Words: MCL	1,088	9,746	160,526	170,272

Table 2: Statistics about the lexicons we built using bootstrapping with PMI and LR. MCL is the manually-crafted lexicon.

5 The System

Below we describe our baseline system: the preprocessing, and the features used.

5.1 Preprocessing

For pre-processing, we applied various algorithms, which we combined in order to achieve better performance. We used Christopher Potts' tokenizer,[10] and we had to be careful since we had to extract not only the words but also other tokens such as hashtags, emoticons, user names, etc. The pre-processing of the tweets goes as follows:

1. **URL and username removal**: tokens such as URLs and usernames (i.e., tokens starting with @) were removed.
2. **Stopword removal**: stopwords were filtered out based on a word list (146 words).
3. **Repeating characters removal**: consecutive character repetitions in a word were removed, e.g., 'какоооо' became 'како' ('what' in English); also were removed repetitions of a word in the same token, e.g., 'дадада' became 'да' ('yes' in English).
4. **Negation handling**: negation was addressed using a predefined list of negation tokens, then the prefix NEG_CONTEXT_ was attached to the following tokens until a clause-level punctuation mark, in order to annotate it as appearing in a negated context, as suggested in (Pang et al., 2002). A list of 45 negative phrases and words was used to signal negation.

[9] We lost some terms in the process of translation, e.g., because some English words translated to the same Macedonian word.

[10] http://sentiment.christopherpotts.net/tokenizing.html

		PMI			LR			PMI + LR
Seeds for bootstrapping	**Source**	**B**	**B+S**	**B+S+M**	**B**	**B+S**	**B+S+M**	**B+S+M**
–	–	–	61.99	78.18	–	61.99	78.18	78.18
1+1 words	(Turney, 2002)	51.48	62.29	78.40	59.82	63.57	78.51	78.89
2+3 smileys	(Mohammad et al., 2013)	61.12	63.81	78.69	65.18	68.99	78.95	79.62
5+5 words	manually-selected	62.55	64.59	79.25	66.70	69.73	80.13	80.87
7+7 words	(Turney and Littman, 2003)	63.02	66.27	79.71	66.98	69.82	80.54	81.99
30+30 words	(Mohammad et al., 2013)	63.47	68.51	79.84	67.28	70.01	80.68	81.33
50+50 words	our MCL	67.11	72.48	80.89	69.79	74.15	81.96	82.73
100+100 words	our MCL	70.94	76.30	82.76	71.41	77.47	84.72	85.45
250+250 words	our MCL	72.25	84.72	92.23	73.76	85.89	93.47	93.55
459+629 words	our MCL	73.82	90.91	94.12	75.29	91.02	94.32	94.44

Table 3: Sentiment classification results (F-score) using lexicons bootstrapped with PMI, LR, or both to calculate $SO(w)$: B = using the bootstrapped lexicon only, B+S = also using non-lexicon features, B+S+M = also using our MCL. The first line shows results when no bootstrapped lexicon is used.

5. **Non-standard to standard word mapping**: non-standard words (slang) were mapped to an appropriate form, according to a manualy crafted predefined list of mappings.

6. **PoS tagging**: rule-based, using a dictionary.

7. **Tagging positive/negative words**: positive and negative words were tagged as `POS` and `NEG`, using sentiment lexicons.

8. **Stemming**: rule-based stemming was performed, which removes or replaces some prefixes and suffixes.

In sum, we started the transformation of an input tweet by converting it to lowercase, followed by removal of URLs and user names. We then normalized some words to Standard Macedonian using a dictionary of 173 known word transformations, and we also removed the stopwords (from a list of 146 words). As part of the transformation, we marked the words in a negated context.

We further created a rule-based stemming algorithm with a list of 65 rules for removing/replacing prefixes and suffixes, inspired by the Porter stemmer (Porter, 1980). We used two groups of rules: 45 rules for affix removal, and 20 rules for affix replacement. Developing a stemmer for Macedonian was challenging as this is a highly inflective language, rich in both inflectional and derivational forms.

5.2 Features

In order to evaluate the impact of the sentiment lexicon, we defined features that are fully or partially dependent on the lexicons. When using multiple lexicons at the same time, there are separate instances of these features for each lexicon. Here are the features we used: number of positive terms, number of negative terms, ratio of the number of positive terms to the number of positive+negative terms, ratio of the number of negative terms to the number of positive+negative terms, sum of all positive scores, sum of all negative scores, sum of all scores, both positive and negative.

For classification, we used logistic regression. Our basic features were TF.IDF-weighted unigrams and bigrams, and also emoticons. We further included additional features that focus on the positive and on the negative terms that occur in the tweet together with their scores in the lexicon. In case of two or more lexicons being used together, we had a copy of each feature for each lexicon.

6 Experiments and Evaluation

Our evaluation setup follows that of the SemEval 2013-2016 task on Sentiment Analysis on Twitter (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov et al., 2016a), and uses an F-score that is the average of the F_1 score for the positive, and the F_1 score for the negative class. Note that, even though implicit, the neutral class still matters in this score. Note also that our focus here is on assessing the impact of our proposed lexicon generation method, and not the classifier itself.

Table 3 shows the results when using each of the bootstrapped lexicons from Table 2. The upper part of the table shows experiments with translations of the seeds used in related work, as described above, while the lower part shows results with (a random subset) of our manually crafted lexicon. We can see that all lexicons outperform the *no bootstrapped lexicons* baseline. The results indicate that our manual lexicon is more useful than the bootstrapping lexicons built using small seeds: it improves over the baseline by 16 points absolute, while the NRC-style or Turney-style lexicons only improve by 2-8 points.

Moreover, using our manually crafted lexicon as a seed for bootstrapping works better than using it as a lexicon: 90.91 vs. 78.18 (with PMI). Moreover, combining it with a bootstrapped lexicon built using all 1,088 words as seeds yields an F-score of 94.12 (with PMI). Note that LR performs better than PMI, by up to four points. Yet, as the last column shows, there is also gain when combining them.

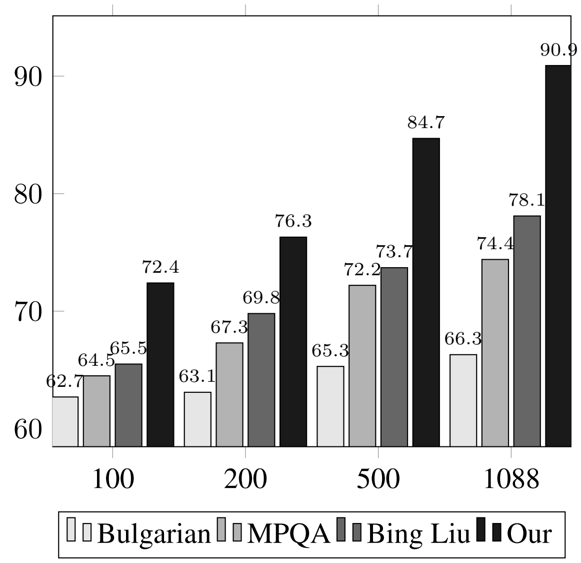

Figure 1: Sentiment polarity classification results (F-score) using different translated bootstrapped lexicons and numbers of seeds with PMI for $SO(w)$, and using B+S as features.

The B+S+M columns in Table 3 show results when using our manually-crafted 1,088-word lexicon as an additional lexicon in each experiment. We can see consistent improvement ranging from 3 to 15 points of F-score absolute on top of the performance of the bootstrapped lexicons. Most interestingly, the bigger the size of the seed, the better the performance of the resulting lexicon (improvement of up to 28 points). So, is it all about the size of the resulting lexicon (as Table 2 shows, bigger seeds yield bigger bootstrapped lexicons)? In order to test this hypothesis, we built bootstrapping lexicons with 100, 200, 500, and 1,088 seeds, with the seeds coming from our lexicons and from the three translated lexicons above. The results are shown in the Figure 1. There are consistent gains as the number of seeds increases, and this is true for seeds coming from our lexicon and also from translations of MPQA, Bing Liu's lexicon, and the Bulgarian lexicon.

However, not all seeds are created equal, even when they are of equal size, and we can see that it is much better to use our manually crafted lexicon as a source of seeds. Yet, if only using 100 seeds from our lexicon, the resulting bootstrapped lexicon would not be able to compete against one built using 1,088 seeds from MPQA or Bing Liu's lexicon. Next, we computed what proportion of each translated lexicon is contained in our lexicon – Bulgarian: 25%, MPQA: 37%, Bing Liu: 43%. We can see that the larger the overlap the better the lexicon, i.e., closer to our domain.

Thus, we can conclude that it is preferable to use (*i*) a *manually-crafted/in-domain lexicon* for the seeds, and (*ii*) a *mid-sized set of seeds*.

7 Conclusion and Future Work

We have presented experiments with different seeds for bootstrapping sentiment polarity lexicons. We have shown that it is best to use (*i*) a *mid-sized seed*, contrary to what is common practice, and (*ii*) a *manually-crafted/in-domain lexicon*, and (*iii*) a classifier such as LR rather than PMI. We have released all our Macedonian lexicons freely for research use.[11]

In future work, we plan experiments for other languages, other sets of seeds, other lexicons, and other learning methods. We further want to study the impact of the raw corpus size, e.g., we could only collect half a million tweets for Macedonian, while Mohammad et al. (2013) used 135 million English tweets. Also, we are interested not only in quantity but also in quality, i.e., in studying the impact of the quality of the individual words when used as seeds. An interesting work in that direction, even though in a different domain and context, is that of (Kozareva and Hovy, 2010).

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC '10, Valletta, Malta.

Daniel Balchev, Yasen Kiprov, Ivan Koychev, and Preslav Nakov. 2016. PMI-cool at SemEval-2016 task 3: Experiments with PMI and goodness polarity lexicons for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, pages 844–850, San Diego, California, USA.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, and Martin Jaggi. 2016. SwissCheese at SemEval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, pages 1124–1128, San Diego, California, USA.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC '06, pages 417–422, Genoa, Italy.

Andrej Gajduk and Ljupco Kocarev. 2014. Opinion mining of text documents written in Macedonian language. *arXiv preprint arXiv:1411.4472*.

Wei Gao and Fabrizio Sebastiani. 2015. Tweet sentiment: From classification to quantification. In *Proceedings of the 7th International Conference on Advances in Social Network Analysis and Mining*, ASONAM '15, pages 97–104, Paris, France.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 470–478, Denver, Colorado, USA.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, Seattle, Washington, USA.

Dame Jovanoski, Veno Pachovski, and Preslav Nakov. 2015. Sentiment analysis in Twitter for Macedonian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '15, pages 249–257, Hissar, Bulgaria.

Borislav Kapukaranov and Preslav Nakov. 2015. Fine-grained sentiment analysis for movie reviews in Bulgarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 266–274, Hissar, Bulgaria.

[11] `http://github.com/badc0re/sent-lex`

Svetlana Kiritchenko, Saif M Mohammad, and Mohammad Salameh. 2016. SemEval-2016 task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, pages 42–51, San Diego, California, USA.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the International Conference on Weblogs and Social Media*, ICWSM '11, Barcelona, Spain.

Zornitsa Kozareva and Eduard Hovy. 2010. Not all seeds are equal: Measuring the quality of text mining seeds. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT '10, pages 618–626, Los Angeles, California, USA.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–74, 3.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463. Springer.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, VarDial '16, Osaka, Japan.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises*, SemEval '13, pages 321–327, Atlanta, Georgia, USA.

Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, pages 31–41, San Diego, California, USA.

Saif Mohammad. 2012. #Emotional tweets. In *Proceedings of *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task*, *SEM '12, pages 246–255, Montreal, Canada.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 312–320, Atlanta, Georgia, USA.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016a. SemEval-2016 Task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, pages 1–18, San Diego, California, USA.

Preslav Nakov, Sara Rosenthal, Svetlana Kiritchenko, Saif M. Mohammad, Zornitsa Kozareva, Alan Ritter, Veselin Stoyanov, and Xiaodan Zhu. 2016b. Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation*, 50(1):35–65.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL '05, pages 115–124, Ann Arbor, Michigan, USA.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '02, pages 79–86, Philadelphia, Pennsylvania, USA.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

Maria Pontiki, Harris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 27–35, Dublin, Ireland.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '2015, pages 486–495, Denver, Colorado, USA.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Veselin Raychev and Preslav Nakov. 2009. Language-independent sentiment analysis using subjectivity and positional information. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '09, pages 360–364, Borovets, Bulgaria.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 73–80, Dublin, Ireland.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 450–462, Denver, Colorado, USA.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. SemEval-2015 task 9: CLIPEval implicit polarity of events. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 442–449, Denver, Colorado, USA.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, March.

Aliaksei Severyn and Alessandro Moschitti. 2015. On the automatic learning of sentiment lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '15, pages 1397–1402, Denver, Colorado, USA.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Veselin Stoyanov and Claire Cardie. 2008. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics*, COLING '08, pages 817–824, Manchester, United Kingdom.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '07, pages 70–74, Prague, Czech Republic.

Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 417–424, Philadelphia, Pennsylvania, USA.

Vasilija Uzunova and Andrea Kulakov. 2015. Sentiment analysis of movie reviews written in Macedonian language. In *ICT Innovations 2014*, pages 279–288. Springer.

Julio Villena-Román, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González Cristóbal. 2013. TASS - Workshop on Sentiment Analysis at SEPLN. *Procesamiento del Lenguaje Natural*, 50:37–44.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Comput. Linguist.*, 30(3):277–308, September.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT-EMNLP '05, pages 347–354, Vancouver, British Columbia, Canada.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, pages 1–9, Hissar, Bulgaria.

Xiaodan Zhu, Svetlana Kiritchenko, and Saif M. Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '14, pages 437–442, Dublin, Ireland.