# Fine-Grained Sentiment Analysis for Movie Reviews in Bulgarian

**Borislav Kapukaranov**
Faculty of Mathematics and Informatics
Sofia University
Bulgaria
b.kapukaranov@gmail.com

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

## Abstract

We present a system for fine-grained sentiment analysis in Bulgarian movie reviews. As this is pioneering work for this combination of language and sentiment granularity, we create suitable, freely available resources: a dataset of movie reviews with fine-grained scores, and a sentiment polarity lexicon. We further compare experimentally the performance of classification, regression and ordinal regression in a 3-way, 5-way and 11-way classification setups, using as features not only the text from the reviews, but also contextual information in the form of metadata, e.g., movie length, director, actors, genre, country, and various scores: IMDB, Cinexio, and user-average. The results show that adding contextual information yields strong performance gains.

## 1 Introduction

With the recent explosion in the popularity of Web forums and social media, sentiment analysis has emerged as a hot research topic. As sentiment-annotated data became readily available, researchers tapped into it and started developing various models for sentiment polarity prediction. Nowadays, there are many applications for sentiment analysis, e.g., businesses getting automatically classified feedback from customers, automated review scoring in retail Web sites, exploration of positive and negative trends, etc.

Movie reviews are a popular and widely available source of sentiment-annotated data. Unlike reviews produced by critics, those contributed by users are typically short and serve primarily to provide brief justification of a user's rating. An important characteristic of movie reviews compared to other sentiment sources is that they are commonly scored on a 5-star scale.

This is very different from sentiment analysis on Twitter, where three-way sentiment classification schemes (*positive, negative, neutral*) have been preferred, e.g., at SemEval 2013-2015 (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov et al., 2015). In contrast, the star system makes the task more fine-grained, thus allowing to capture user opinion better.

Here, we present experiments in predicting fine-grained stars, including halves, for Bulgarian movie reviews. This is a challenging task, that can be seen as (a) *multi-way classification*, i.e., choosing one out of eleven classes, (b) *regression*, i.e., predicting a real number, or (c) something in between, namely *ordinal regression*, i.e., predicting eleven values, but taking ordering into account, e.g., predicting 4 when the actual value is 3.5 would be better than predicting 1.

While sentiment classification in movie reviews has been extensively studied for English (movie reviews datasets were among the earliest to use for this task), it has not been tried for Bulgarian so far. Moreover, most research has focused on positive/negative/neutral classification and finer-grained schemes have been less popular (as they are harder). Even when used, the focus has typically been on having just five categories, not allowing halves. Thus, our contributions in this paper can be summarised as follows:

- We create a new dataset for movies in Bulgarian,[1] where each review is associated with an 11-scale star rating: 0, 0.5, 1, ..., 4.5, 5.
- We prepare a new sentiment lexicon for Bulgarian, which is also freely available.
- Most importantly, we present the first work for Bulgarian on predicting fine-grained sentiment.

[1] The dataset is freely available for research purposes at http://bkapukaranov.github.io/

*Proceedings of Recent Advances in Natural Language Processing*, pages 266–274,
Hissar, Bulgaria, Sep 7–9 2015.

The remainder of this paper is organized as follows: Section 2 introduces related work, Section 3 describes the dataset and teh lexicon we prepared, Section 4 presents the features we experiment with, Section 5 describes our experiments, and Section 6 discusses the results. Finally, Section 7 concludes and points to some possible directions for future work.

## 2 Related Work

Pang et al. (2002) were the first to look into text classification not in terms of topics, but focusing on how sentiment polarity is distributed in a document. They tried several machine learning algorithms on an English movie reviews dataset, and evaluated the performance of basic features such as $n$-grams and part of speech (POS) tags.

Movie reviews were one of the first research domains for sentiment analysis as they (*i*) have the properties of a short message, and (*ii*) are already manually annotated by the author, as the score generally reflects sentiment polarity. Popular features for score/sentiment prediction include POS tags, word $n$-grams, word lemmata, and various context features based on the distance from a topic word. The challenge with movie reviews is that only some of the words are relevant for sentiment analysis. In fact, often the review is just a short narrative of the movie plot. One way to approach the problem is to use a subjectivity classifier (Pang and Lee, 2004), which can be used to filter out objective sentences from the reviews, thus allowing the classifier then to focus on the subjective sentences only.

Early 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 looked into the linguistic aspects of how opinions, evaluations, and speculations are expressed in text (Wiebe et al., 2004), into the role of context for 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 was crucial to know not just the sentiment, but also the topic towards which this sentiment was expressed (Stoyanov and Cardie, 2008).

Fine-grained sentiment analysis tries to predict sentiment in a text using a finer scale, e.g., 5-stars; Pang and Lee (2005) pioneered this sub-field. In their work, they looked at the problem from two perspectives: as one vs. all classification, and as a regression by putting the 5-star ratings on a metric scale. An interesting observation in their research is that humans are not very good at doing such kinds of highly granular judgments and are often off the target mark by a full star.

Naturally, most research in sentiment analysis was done for English, and very little efforts were devoted to other languages. We are not aware of other work on fine-grained sentiment analysis for Bulgarian. There is work on sentiment analysis by Bulgarian scolars (Raychev, 2009; Raychev and Nakov, 2009; Kraychev and Koychev, 2012; Kraychev, 2014).

We are aware of three publications for the closely-related Macedonian language,[2] which is mutually intelligible with Bulgarian.

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

More closely related to our work, Uzunova and Kulakov (2015) experimented with 400 movie reviews (200 positive + 200 negative), 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.

Finally, Jovanoski et al. (2015) presented work on sentiment analysis of Macedonian tweets (8,583 for training + 1,139 for testing) using a 3-way tweet-level sentiment polarity classification scheme: positive, negative, and neutral/objective. They used standard features but variety of preprocessing steps, including morphological processing and POS tagging for Macedonian, negation handling, text standardization, tweet-specific processing, etc. More imporantly, they made use of several lexicons, some translated from other languages,[3] which they augmented with bootstrapping, ultimately achieving results that are on par with the state of the art for English.

[2] Some linguists consider Macedonian a dialect of Bulgarian; this is also the position of the Bulgarian government.

[3] In fact, they used, without translation, the Bulgarian lexicon that we present in this work.

Given the lack of previously developed datasets or sentiment polarity lexicons for Bulgarian, we had to create them ourselves. In addition to preparing a dataset of annotated movies, we further focused on building a sentiment polarity lexicon for Bulgarian. This is because lexicons are crucial for sentiment analysis. Since the very beginning, researchers have realized that sentiment analysis was quite different from standard document classification (Sebastiani, 2002), e.g., into categories such as *business*, *sport*, and *politics*, and that sentiment analysis crucially needed external knowledge in the form of suitable sentiment polarity lexicons. For further detail, see the surveys by Pang and Lee (2008) and Liu and Zhang (2012).

Until recently, such sentiment polarity lexicons were manually crafted, and were thus of small to moderate size, e.g., LIWC (Pennebaker et al., 2001), General Inquirer (Stone et al., 1966), Bing Liu's lexicon (Hu and Liu, 2004), and MPQA (Wilson et al., 2005), all have 2000-8000 words. Early efforts in building them automatically also yielded lexicons of moderate sizes (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 2013-2015 (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015). These lexicons were crucial for the top-performing teams in the competition in all three years.

Similar observations were made in the Aspect-Based Sentiment Analysis task at SemEval 2014-2015 (Pontiki et al., 2014). In both tasks, the winning systems benefited from building and using massive sentiment polarity lexicons (Mohammad et al., 2013; Zhu et al., 2014).

## 3 Data

Our dataset consist of 347 movies with a total of 10,198 Bulgarian reviews, which we crawled from the ticket-booking website Cinexio.[4] We chose only movies for which scored reviews in Bulgarian were present on the website. For each movie, we include a set of user reviews, each annotated with a score on an 11-point scale: 0, 0.5, 1, ..., 4.5, 5 stars. More detailed statistics about our movie reviews dataset can be found in Table 1.

| Characteristic | Count |
|---|---|
| unique words | 8,406 |
| unique users | 3,395 |
| unique movie genres | 23 |
| unique movie countries | 49 |
| unique movie actors | 1,668 |
| unique movie directors | 317 |

Table 1: Statistics about our dataset.

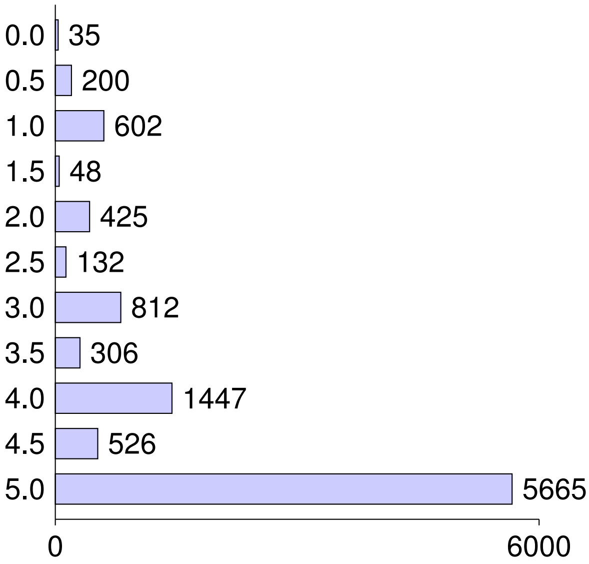

Figure 1: User rating distribution in our dataset.

Figure 1 shows a distribution of the user ratings in our movie reviews dataset. We can see that the distribution is generally skewed towards full scores, while scores with halves are much less frequent: people seem to prefer a 5-point scale, and would not take full advantage of an 11-point one. Moreover, the distribution is also skewed towards high scores, and quite heavily towards a 5-star rating in particular.

In addition to the movie reviews dataset, we further automatically generated a sentiment polarity lexicon for Bulgarian, using that dataset and pointwise mutual information (PMI) with respect to the positive and to the negative class, following the idea presented in (Turney, 2002):

$$pmi(w, class) = \log \left[ \frac{p(w \,\&\, class)}{p(w)\, p(class)} \right] \quad (1)$$

Then, we calculated a sentiment polarity score:

$$polarity = pmi(w,\, pos) - pmi(w,\, neg) \quad (2)$$

Words with high positive/negative polarity were included in our sentiment polarity lexicon; this included 5,016 positive and 2,415 negative words.

[4] http://www.cinexio.com

Here, we list some examples of movie reviews (with English translations in footnotes):

- "добре, че го бастисаха накрая, че да се спре с тая пропаганда.. ;)"[5] with score 3.5

  This example is very interesting as it expresses cheerful and light mood, as indicated by the winking emoticon. Yet, it also mentions *propaganda*, which hints slight irritation about the way the movie plot developed.

- "Много добър филм"[6] with score 5.0

  Nothing really surprising here: typical positive comment, without going into specifics.

- "Много добър филм. Просто ни е далечен Американския патриотизъм."[7] with score 4.0

  Here, despite having the same text as the max-scored previous example, the review author has given a slightly lower score. From the text alone, we can conclude that the author likes the movie very much, but still makes a general remark on how the movie will be accepted culturally by Bulgarian viewers.

- "Не ме впечатли."[8] with score 3.0

  A typical mid-score comment: direct, clear.

- "Доста тъжно :("[9] with score 5.0

  This is a perfect example of why this problem is hard. The text alone shows clearly negative emotions both indicated by text and by the crying emoticon, but it seems that the author actually liked the movie very much and gave it a maximum score.

Negative reviews from other movies:

- "доста дълъг и поне да се случваше нещо..."[10] with 1.5 score

  On the low end of the scale, the scores become highly subjective, and often the same wording can be annotated with a full star difference in the score.

- "Филма е само част от трилогия, помнете че историята свършва найинтересното"[11] with score 2.0

  This is another confusing example. Did the author actually like the movie? Or was s/he affected by somebody else's opinion?

- "Доста повече екшън от първата част"[12] with score 3.0

  This is a great example showing that the perception of movies in a multipart series is influenced by earlier parts. It is not clear what people are scoring: the entire series or just the current (latest) part of the movie? People naturally try to compare with earlier series, which influences their scores.

In general, scores could be heavily biased, and also relative: if one has recently watched a bad movie, the following movie, even if just slightly better, could get an inflated score.

## 4 Features

In this section, we describe the features we experimented with: textual and contextual.

### 4.1 Textual Features

We used the following textual features:

- **words:** binary feature for each word;
- **emoticons:** binary feature for each positive/negative emoticon;
- $n$**-grams:** binary feature for each $n$-gram (we only used bigrams).
- **lexicon:** We further included two features based on our automatically generated movie reviews lexicon. They represent the positive and the negative overall score of the movie review, obtained by aggregating the lexicon scores of each word in the review text.

Note that our dataset lacks enough relevant instances to use features such as all-caps and punctuation, and thus we did not use them here.

Moreover, we found that using bigram features did not make much difference for this particular dataset, therefore the final feature set for the baseline system only used *bag of words*, *emoticons*, and the *lexicon* features.

[5] "it is good that they got him in the end, so the propaganda could finally be over.. ;)"

[6] "Very good movie"

[7] "Very good movie, we are just a little bit off on the American partriotic message"

[8] "Not impressed"

[9] "Quite sad :("

[10] "quite long, on top of that nothing actually happens..."

[11] "The movie is just the first part of a series, keep in mind the story ends in the most interesting part"

[12] "Definitely more action compared to the first part"

### 4.2 Contextual Features

In addition to the above textual features, we further added some contextual (metadata) features:

- **movie length**: numeric feature indicating the run-length of the movie;
- **country**: binary feature indicating the country the movie comes from;
- **genres**: indicator feature for each genre;
- **actors**: indicator feature for each actor;
- **director**: indicator feature for each director;
- **average user rating**: numeric feature with the user's average movie review score;
- **IMDB score**: numeric feature, current average score for this movie in IMDB;
- **Cinexio score**: numeric feature, current average score for this movie in Cinexio.

## 5 Experiments and Evaluation

Below we describe the class granularities we experimented with, the learning algorithms we used, and the evaluation results.

### 5.1 Class Granularity

In the original formulation, we have eleven classes: 0, 0.5, 1, ..., 4.5, 5. For model comparison purposes, we further experimented with aggregated classes. Thus, we ended up with three class inventories of various sizes:

- **11-way**: includes all labels, both integer and half-star;
- **5-way**: includes only the full stars;
- **3-way**: divides the scores into three classes, $positive \geq 3.5 > neutral \geq 2 > negative.$

### 5.2 Learning Algorthms

We performed experiments with three machine learning approaches: (*i*) classification, (*ii*) regression, and (*iii*) ordinal regression. We evaluated using a 5-fold cross validation. For scoring, we used the same metric for all class inventories and for all learning approaches, namely Mean Squared Error (MSE), which is standard for a task asking to predict ordinal values as in our case.

**Classification.** For classification, we used SVM with a linear kernel and L2-regularized L2-loss, as implemented in LibSVM (Chang and Lin, 2011). We used a one vs. all model, which we applied for each class inventory size: 3, 5, 11.

**Regression.** For regression, we used the same SVM tool and the same features and parameters as for classification, but we predicted a numerical value; this is known as *support vector regression* (Smola and Schölkopf, 2004)

**Ordinal Regression.** For this scenario, we used *ordinal logistic regression*. This model is also known as *proportional odds* and was introduced by McCullagh (1980).[13] The use of ordinal regression for sentiment analysis, is not very common, mostly because the ordinal formulation of the task is not very common, even though it was used by some researchers (Pang and Lee, 2005; Goldberg and Zhu, 2006; Baccianella et al., 2009). Yet, it makes a lot of sense to use it as it tries to fit the data into thresholded regions as a classification task would do, and at the same time tries to predict values with an established order and position in the label space. This makes it interesting especially in the 5-class setup, where we have a small number of labels and there is ordering between them.

### 5.3 Results

Our preliminary cross-validation experiments have shown that not all features that we have introduced above were really relevant; thus, we created a selected set of highly-relevant features: *words*, *emoticons*, *lexicons*, *Cinexio score*, and *average user rating*. We used this feature set when comparing the three machine learning algorithms (classification, regression, and ordinal regression), for the three class sizes (3, 5, and 11). The results are shown in Table 2.

| **Model** | **11-way** | **5-way** | **3-way** |
|---|---|---|---|
| Classification | 1.041 | 0.666 | 0.141 |
| Regression | 0.484 | 0.472 | 0.135 |
| Ordinal regression | 1.438 | 1.276 | 0.464 |

Table 2: **Evaluation using the selected features.** Shown is MSE for the three machine learning algorithms and for the three class sizes. (Lower scores are better.)

[13] There are several alternative machine learning approaches to ordinal regression, e.g., *support vector ordinal regression* (Chu and Keerthi, 2007).

| Feature | MSE | ΔMSE |
|---|---|---|
| *baseline* *(all textual features)* | *0.745* | – |
| bl + IMDB score | 0.689 | -0.056 |
| bl + Cinexio score | 0.669 | -0.076 |
| bl + Cinexio + IMDB | 0.658 | -0.087 |
| bl + user avg. score | 0.520 | -0.225 |
| bl + user avg. score + Cinexio | 0.484 | -0.261 |
| bl + movie length | 0.484 | -0.261 |
| bl + director | 0.732 | -0.013 |
| bl + country | 0.723 | -0.022 |
| bl + actors | 0.484 | -0.261 |
| bl + genres | 0.723 | -0.022 |

Table 3: **Impact of individual contextual features when added to the baseline.** Shown is MSE for the regression model with 11 classes.

| Feature | MSE | ΔMSE |
|---|---|---|
| *all* *(all textual + contextual features)* | *0.515* | – |
| all − words | 0.523 | +0.008 |
| all − lexicons | 0.745 | +0.230 |
| all − emoticons | 0.515 | 0.000 |
| all − IMDB score | 0.494 | -0.021 |
| all − Cinexio score | 0.544 | +0.029 |
| all − user avg. score | 0.736 | +0.221 |
| all − movie length | 0.515 | 0.000 |
| all − directors | 0.515 | 0.000 |
| all − country | 0.514 | -0.001 |
| all − actors | 0.515 | 0.000 |
| all − genres | 0.514 | -0.001 |

Table 4: **Impact of individual features when excluded from the full feature set.** Shown is MSE for the regression model with 11 classes.

We can see in Table 2 that the best results are achieved for *regression*, where the mean squared error is within half a point away for the 11-way and the 5-way class inventories, and it is about four times lower for the 3-way one. The second-best performing machine learning approach is *classification*; its performance is very close to that of regression on the 3-way class inventory, but the gap widens with 5 classes (about 50% difference), and becomes huge with 11 classes (100% difference), where the predictions are on average a full point off from the target. Finally comes the worst-performing approach, *ordinal regression*, which consistently performs about four times worse than the standard regression.

Interestingly, while *classification* performs badly compared to *regression* on the 11-way class inventory, it quickly catches up for smaller numbers of classes, and the two learning approaches get quite close on the 3-way class inventory. This is expected, as classification usually struggles with too many class labels, especially in the case of uneven class distribution, and this is indeed our case, as we have seen in Figure 1.

However, the low performance of ordinal regression is quite surprising; the expectation was that it would perform the best. In future work, we plan to have a closer look at the reasons for these results. At this point, we can only note that we used SVM as the basic underlying classifier in our *classification* and *regression* experiments, but we used *logistic regression* as the basis for our *ordinal regression*. It is unclear whether this alone could explain the difference in performance, though.

Table 3, shows the impact of the individual context features (and some feature combinations) when added to the baseline textual features. We report results for 11-way classification with the *regression* model; and the last column shows the difference in MSE compared to the baseline. We can see that each of the features yields improvements, which means that they all are indeed relevant. The most important features turn out to be *movie length*, *actors*, and *user average score*.

Yet, some features might be redundant, i.e., having one feature might mean that we do not need to have some other ones. In order to study this, we performed experiments excluding features one at a time from the full set of features, both textual and contextual. The results are shown in Table 4. As before, we study 11-way classification with the *regression* model. The relative change in MSE compared to the full model is shown in the last column of the table. We can see that *lexicons* have the biggest impact, which is to be expected, as we know from previous work that they are among the most important resources for sentiment analysis. Another strong feature turns out to be the *user average score*, which also makes sense: a user who has been giving high scores in the past is likely to give high scores in the future. We can further see that many contextual features, e.g., *movie length, actors, director, genres* and *country*, made almost no difference. This is surprising as the first two yielded the largest improvements over the *baseline* features in Table 3; we believe this reflects feature interaction, but we plan closer investigation.

## 6 Discussion

We have seen in our experiments above that the best-performing model used *regression* and *contextual* features, in addition to *textual* ones. We believe that the kind of context we model, primarily metadata, is indeed important as, while it is not present in the text of the review, it has been taken into account when the author rated the movie.

Interestingly, we have found that factual information was not very useful. This is a good sign as it suggests that Cinexio users seem not to have prejudice about the expected quality of a movie based on its country of origin, director(s), or genre; however, actors playing do have impact.

One of the most useful contextual features was the *user average score*. Some users tend to give consistently high/low scores regardless of the movie, and thus knowing their average scores allows us to take this into account.

A related useful feature was the *Cinexio score* of the target movie. The idea is that if a movie has a high/low overall score, we should expect a new user also to give it a high/low score. While *IMDB scores* are quite similar, we had mixed results for them: they were quite helpful compared to the baseline, but were harmful with respect to the full set of features.

Given the difference between Cinexio and IMDB scores, we decided to have a closer look at how they relate to each other. This is shown in Figure 2. The blue line connects the corresponding Cinexio–IMDB scores, while the red line shows how perfect correlation would look like. Note that IMDB scores are in the 0–10 range.

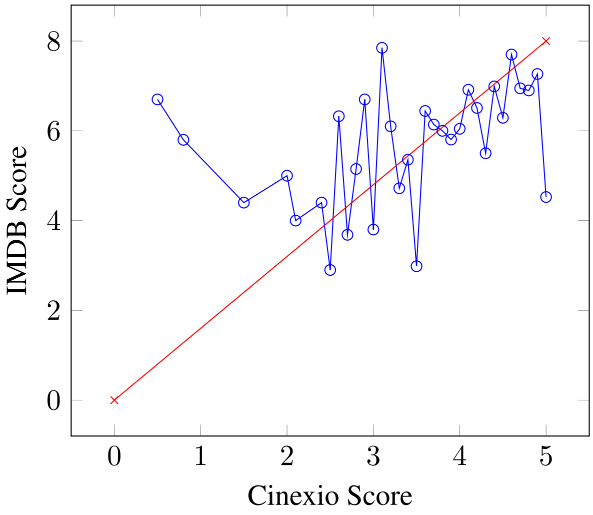

Figure 2: Cinexio vs. IMDB scores.

This is an interesting plot as it reflects how viewers (a) in Bulgaria and (b) worldwide feel about the same movie. We can see that the general correlation is there, especially for the mid-high scores. However, there is a lot of discrepancy with the extreme scores, i.e., what Bulgarian viewers see as extremely good is regarded as average at IMDB, and what they consider extremely bad, actually has an above-average score at IMDB.

This discrepancy in IMDB vs. Cinexio scores explains the mixed results we got when using the IMDB score as a feature. One way to fix this could be to split the IMDB feature into several features, each responsible for just a sub-interval of the possible values of the original feature. This might be useful for some other features with numerical values, which could show non-linearity, e.g., *Cinexio score*, *average user score*, or *movie length*.

## 7 Conclusion and Future Work

We presented the first research on fine-grained sentiment analysis for Bulgarian. As this is pioneering work for this language, we created a suitable dataset and a sentiment polarity lexicon, which we made freely available for research purposes; this should enable further research.

We further compared experimentally the performance of classification, regression and ordinal regression in a 3-way, 5-way and 11-way classification setups, using as features not only the text from the reviews, but also contextual information in the form of metadata, e.g., movie length, director, actors, genre, country, and various scores: IMDB, Cinexio, and user-average. The experimental results have shown that adding contextual information yields strong performance gains.

In future work, we plan to investigate the low performance of ordinal regression. We further want to experiment with more features, e.g., summary of the plot, subtitles, information from other websites such as IMDB, as well as with more linguistic processing of the text, e.g., stemming (Nakov, 2003b; Nakov, 2003a), POS tagging (Georgiev et al., 2012), and named entity recognition (Georgiev et al., 2009). We also want to see the impact of earlier comments on the sentiment of newer comments (Vanzo et al., 2014; Barrón-Cedeño et al., 2015; Joty et al., 2015). Finally, we would like to apply our system to help other tasks, e.g., finding trolls in Web forums (Mihaylov et al., 2015a; Mihaylov et al., 2015b).

## References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Multi-facet rating of product reviews. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 461–472, Toulouse, France.

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.

Alberto Barrón-Cedeño, Simone Filice, Giovanni Da San Martino, Shafiq Joty, Lluís Màrquez, Preslav Nakov, and Alessandro Moschitti. 2015. Thread-level information for comment classification in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '15, pages 687–693, Beijing, China.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Wei Chu and S. Sathiya Keerthi. 2007. Support vector ordinal regression. *Neural Comput.*, 19(3):792–815, March.

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*.

Georgi Georgiev, Preslav Nakov, Kuzman Ganchev, Petya Osenova, and Kiril Simov. 2009. Feature-rich named entity recognition for Bulgarian using conditional random fields. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '09, pages 113–117, Borovets, Bulgaria.

Georgi Georgiev, Valentin Zhikov, Petya Osenova, Kiril Simov, and Preslav Nakov. 2012. Feature-rich part-of-speech tagging for morphologically complex languages: Application to Bulgarian. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 492–502, Avignon, France.

Andrew B. Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs '06, pages 45–52, Sydney, Australia.

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, WA, USA.

Shafiq Joty, Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2015. Global thread-level inference for comment classification in community question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, Lisbon, Portugal.

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

Boris Kraychev and Ivan Koychev. 2012. Computationally effective algorithm for information extraction and online review mining. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12, pages 64:1–64:4, Craiova, Romania.

Boris Kraychev. 2014. *Extraction and Analysis of Opinions and Sentiments from Online Text*. Ph.D. thesis, Sofia University, January. (in Bulgarian).

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.

Peter McCullagh. 1980. Regression models for ordinal data. *J. Roy. Statist. Soc. B*, 42:109–142.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015a. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 310–314, Beijing, China.

Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015b. Exposing paid opinion manipulation trolls. In *Proceedings of the Conference on Computational Natural Language Learning*, Hissar, Bulgaria.

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, GA, USA.

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, GA, USA.

Preslav Nakov, Sara Rosenthal, Svetlana Kiritchenko, Saif Mohammad, Zornitsa Kozareva, Alan Ritter, Veselin Stoyanov, and Xiaodan Zhu. 2015. Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation*.

Preslav Nakov. 2003a. Building an inflectional stemmer for Bulgarian. In *Proceedings of the 4th International Conference on Computer Systems and Technologies*, CompSysTech '03, pages 419–424, Sofia, Bulgaria.

Preslav Nakov. 2003b. BulStem: Design and evaluation of an inflectional stemmer for Bulgarian. In *Proceedings of the Workshop on Balkan Language Resources and Tools*, Thessaloniki, Greece.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, pages 271–278, Barcelona, Spain.

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, MI, 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, PA, 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.

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.

Veselin Raychev. 2009. *Automatic Recognition of Positive/Negative Subjectivity in Text*. Ph.D. thesis, Sofia University, March. (in Bulgarian).

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, CO, USA.

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

Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August.

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.

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, PA.

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

Andrea Vanzo, Danilo Croce, and Roberto Basili. 2014. A context-based model for sentiment analysis in Twitter. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pages 2345–2354, Dublin, Ireland.

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, BC, Canada.

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