Stierlitz Meets SVM: Humor Detection in Russian*

Anton Ermilov[1], Natasha Murashkina[1], Valeria Goryacheva[2], and Pavel Braslavski[3,4,1]

[1] National Research University Higher School of Economics, Saint Petersburg, Russia
[2] ITMO University, Saint Petersburg, Russia
[3] Ural Federal University, Yekaterinburg, Russia
[4] JetBrains Research, Saint Petersburg, Russia
{anton.yermilov, murnatty, gor.ler177}@gmail.com, pbras@yandex.ru

Abstract. In this paper, we investigate the problem of the humor detection for Russian language. For experiments, we used a large collection of jokes from social media and a contrast collection of non-funny sentences, as well as a small collection of puns. We implemented a large set of features and trained several SVM classifiers. The results are promising and establish a baseline for further research in this direction.

Keywords: humor recognition · evaluation

1 Introduction

Humor is an important aspect of human communication. Rapid proliferation of conversational agents, voice interfaces, and chatbots, as well as the need to analyze large volumes of social media texts make the task of humor detection highly relevant.

In this study, we used a subset of an existing collection of short jokes in Russian from social media and also collected a contrast collection of non-funny sentences. In addition, we collected a small collection of puns to test the developed method on this special kind of humorous content. We engineered a wide range of features that reflects different aspects of language – lexical, semantic, structural, etc. We trained several binary classifiers and evaluated contribution of individual feature groups to the classification quality. The obtained results demonstrate acceptable performance and provide the basis for further research in this direction. To the best of our knowledge, current study is the first experiment on automatic detection of humor in the Russian language.

* Stierlitz is a Soviet spy working deep undercover in Nazi Germany, a protagonist of a TV series from 1972 based on a novel by Yulian Semionov. Stierlitz became a popular joke character in Soviet and post-Soviet culture.

2 Related Work

The humor recognition is usually formulated as a classification task with a wide variety of features – syntactic parsing, alliteration and rhyme, antonymy and other WordNet relations, dictionaries of slang and sexually explicit words, polarity and subjectivity lexicons, distances between words in terms of *word2vec* representations, etc. In their pioneering work, Michalcea and Strapparava [7] compiled a dataset of humorous and non-humorous sentences in English – 16,000 one-line jokes from the web and 16,000 sentences from the news, the British National Corpus, collections of proverbs, as well as collection of common sense sentences and performed a classification experiment with different features. A follow-up study [6] investigated humor features in more detail. Zhang and Liu [14] experimented with the humor detection in tweets. Yang et al. [13] introduced the notion of humor anchors – words and phrases 'responsible' for a humorous effect, experimented with a large collection of puns and explored a wide range of features for the humor detection, including those based on vector representations. Shahaf et al. [12] addressed the task of ranking cartoon captions provided by the readers of New Yorker magazine. They employed a wide range of linguistic features as well as features from manually crafted textual descriptions of the cartoons. Two recent shared tasks dealing with humor within the SemEval campaign signal a growing interest in the topic [8,9]. A cognate task is detection of other forms of figurative language such as irony and sarcasm [11,10].

3 Data

In the current study we used a collection of jokes in Russian from online social networks that we obtained from the authors of [2]. The collection consists of about 63,000 one-liners collected from VK and Twitter. The jokes are in plain text, i.e. media content, URLs, and hashtags are removed; more details about the dataset can be found in the paper. From this collection, we randomly sampled 47,000 items for our experiments. To build a contrast collection, we gathered sentences from Russian classical novels (28,000), news headlines (13,000) and proverbs (6,000). We did not make efforts to ensure lexical similarity of the funny and non-funny parts of the collection, as the authors of [7] did. The only additional parameter was the length – sentences of 25 words and shorter are included in the collection (average length is 14 words). For experiments, the collection was splitted into training/test sample in a ratio of 80/20.

In addition, we manually created a small collection of puns. In total, there are 200 jokes with a word play in the collection, most of them are associated with the "Omsk Ptitsa" meme and the Stierlitz jokes. We used this collection only for testing classifiers trained on the data from the BIG collection.

4 Features

Based on literature review and manual inspection of the collection, we implemented six groups of text features that can potentially distinguish between humorous and non-humorous content. The features are briefly described below.

Bag-of-words (BOW). Each text is presented as a 12,000-dimensional binary vector. The intuition behind the feature is that some words are quite specific for the humorous content.

Sentence2Vec (S2V) is aimed at capturing sense of the text as a 300-dimensional vector. We summed up vectors of individual words in the text weighed by their IDFs. We used pre-trained word2vec vectors available through the RusVectōrēs project [5]. IDF weights are calculated using the Russian National Corpus data.[5]

Structural features (SF) are shallow features capturing the complexity of the text (average word length in characters and syllables, fraction of stopwords) and its organization – punctuation marks, question words and certain conjunctions.

Lexical features (LF). This group of word-level features includes:

- minimum/maximum word frequencies calculated using RNC statistics;
- a share of words with non-common usage labels (*informal*, *offensive*, *vulgar*, etc.) from the Russian Wiktionary[6];
- a maximum number of possible POS tags over all words and a proportion of nouns/verbs/adjectives/numerals in the text based on the PyMorphy output [4];
- a presence of proper names and parenthetical words.

RuWordNet features (RWN). Using the RuWordNet thesaurus[7] we calculated the following features:

1. An ambiguity
   - a *sense combination*, formalized as $\sum \log(n_{w_i})$, where n_{w_i} is the number of senses of the word w_i (we account only for nouns, verbs and adjectives present in the RuWordNet);
   - the largest *path similarity* over all word-sense pairs, whereas the *path similarity* is the minimal distance between word-senses in thesaurus graph (lower values correspond to semantically closer senses);
2. Domains
   - a number of different domains associated with words in the text;
   - a number of words that belong to different domains.
3. A number of synonym and antonym pairs in the text.

[5] `http://ruscorpora.ru/corpora-freq.html`
[6] `https://ru.wiktionary.org/`
[7] `http://ruwordnet.ru/`

Word2Vec (W2V). Following [13], we calculate two word2vec-based features:

- *disconnection*: the maximum semantic distance of word pairs in a sentence;
- *repetition*: the minimum semantic distance of word pairs in a sentence.

5 Results and Discussion

We used the LibSVM [3] to train classifiers. We experimented with various combinations of feature groups. The Table 1 below summarizes results. The reported figures correspond to the linear SVM that delivered better results in our experiments than SVMs with polynomial and RBF kernels. Columns 2–5 report results achieved on the test set of the 'big' dataset of one-liners and non-funny sentences; precision, recall, and F1 correspond to the humorous class. The last column of the Table reports recall of the classifier trained on the training set from the 'big' dataset and then applied to the small collection of puns.

As can be seen from the Table below, the classification based solely on bag-of-words features is a very strong baseline ($F1 = 0.846$ on the BIG dataset, $R = 0.671$ on the PUNS). On the one hand, it can be explained through the way the collection was built: positive and negative classes are quite distinctive on the lexical level. On the other hand, recall on the independent PUNS collection is also relative high. *S2V* is a runner-up among individual feature groups ($F1 = 0.811$ on the BIG dataset, $R = 0.601$ on the PUNS). Thus, S2V shows no generalization over individual words. We can hypothesize that vector representation 'flattens' the хорошosentence meaning and doesn't account for possible alternative interpretation, which might be crucial for the humorous content. The combination of these two sentence meaning representations (BOW + S2V) improves over both approaches and achieves the best score on the PUNS collection ($recall = 0.695$). Other feature groups, taken separately, demonstrate much lower performance.

The combination of BOW with features, potentially reflecting semantic relations between words in the sentence (RWN and W2V), delivers mixed results. Adding RWN features improves precision on the humorous class ($P = 0.863$), while W2V degrades overall results on the 'big' collection. One can argue that manually crafted semantic resources are still a viable alternative for general-purpose semantic representations based on neural networks, especially for high-precision results. However, these combinations behave reversely on the PUNS collection. BOW + W2V shows second-best result on the PUNS ($R = 0.676$). Results in the Table 1 support in general the claim that more features mean the better classification quality. The combination of all features delivers best results on the BIG dataset ($F1 = 0.884$). However, the addition of two W2V features has a marginal impact. These results somewhat contradict the feature importance considerations reported in [13]. However, a direct comparison between different datasets in different languages is hardly possible.

A manual inspection of misclassified jokes reveals that the majority of them are unfunny according to our subjective opinion. For example, this item from the jokes collection looks rather like a proverb:

Table 1. Humor recognition results.

	BIG				PUNS
Feature set	Accuracy	Precision	Recall	F1	Recall
BOW	0.848	0.855	0.837	0.846	0.671
S2V	0.813	0.820	0.801	0.811	0.601
SF	0.671	0.658	0.715	0.685	0.385
LF	0.618	0.603	0.690	0.643	0.117
RWN	0.563	0.626	0.311	0.416	0.211
W2V	0.527	0.579	0.196	0.293	0.160
BOW + RWN	0.850	0.863	0.832	0.847	0.638
BOW + LF	0.850	0.856	0.842	0.849	0.559
BOW + SF	0.869	0.872	0.863	0.868	0.568
BOW + W2V	0.846	0.853	0.836	0.845	0.676
BOW + SF + LF + RWN	0.871	0.873	0.862	0.870	0.521
S2V + RWN	0.814	0.824	0.798	0.811	0.592
S2V + LF	0.818	0.826	0.806	0.815	0.526
S2V + SF	0.839	0.848	0.826	0.837	0.498
S2V + W2V	0.814	0.822	0.802	0.812	0.606
S2V + SF + LF + RWN	0.846	0.854	0.834	0.844	0.521
BOW + S2V	0.868	0.873	0.861	0.867	**0.695**
BOW + S2V + SF + LF + RWN	**0.885**	**0.892**	**0.875**	**0.884**	0.620
BOW + S2V + SF + LF + RWN + W2V	**0.885**	**0.892**	**0.876**	**0.884**	0.615

Хочешь идти быстро — иди один. Хочешь уйти далеко — идите вместе.

If you want to go fast, go alone. If you want to go far, go together.

Other false negatives are *referential* jokes that require some world knowledge to comprehend them (see [1] for details). For example, this joke refers to dung beetles rolling balls out of dirt and ball-shaped Raffaello candy:

Жук-навозник на День рождения прикатил жене рафаэлку.

A dung beetle brought his wife a Raffaello as a birthday present.

Considering puns, we hypothesize that the following joke was not recognized because of a very scarce context (the pun plays around two senses of the verb *звонить – to ring/to phone*).

Звонил колокол. Угрожал. // The bell rang. Threatened.

Most false positives are literature excerpts, for example:

Все подняли головы, прислушались, и из леса, в яркий свет костра, выступили две, держащиеся друг за друга, человеческие, странно одетые фигуры. // Everyone lifted their heads, listening closely, and two strangely dressed human figures stood out from the forest into the bright light of the fire, holding each other.

Many incorrectly classified excerpts were rather long. Possibly, many word combinations result in triggering some semantic features. Moreover, sentences from fiction works may contain some figurative language.

6 Conclusion

We prepared data and conducted experiments aimed at the humor detection in short Russian texts. We implemented a wide range of text features and conducted a comparative study of their impact on the classification quality. The obtained results form a strong baseline for future research in the field of a computational humor on Russian language data. Pun collection used in the study is freely available for research.[8] In the future, we plan to employ a more elaborate sampling of negative (non-humorous) examples. In addition, we plan to develop methods and features that better capture a word play; expand the collection of puns and conduct a finer-grained annotation of jokes. In the framework of this study, we haven't investigated several features potentially useful for the humor detection: phonetic and syntactic features, as well as those based on sentiment lexicons. We plan to address these tasks in the future.

Acknowledgments. We thank Valeria Bolotova and Vladislav Blinov for sharing their humor dataset, as well as Natalia Loukachevitch for providing us with the RuWordNet data.

References

1. Attardo, S.: Linguistic theories of humor. Mouton de Gruyter (1994)
2. Bolotova, V., et al.: Which IR model has a better sense of humor? Search over a large collection of jokes. In: Dialogue. pp. 29–42 (2017)
3. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011)
4. Korobov, M.: Morphological analyzer and generator for Russian and Ukrainian languages. In: AIST. pp. 320–332. Springer (2015)
5. Kutuzov, A., Kuzmenko, E.: Webvectors: A toolkit for building web interfaces for vector semantic models. In: AIST. pp. 155–161 (2017)
6. Mihalcea, R., Pulman, S.: Characterizing humour: An exploration of features in humorous texts. In: CICLing. pp. 337–347 (2007)
7. Mihalcea, R., Strapparava, C.: Learning to laugh (automatically): Computational models for humor recognition. Computational Intelligence **22**(2), 126–142 (2006)
8. Miller, T., Hempelmann, C., Gurevych, I.: SemEval-2017 Task 7: Detection and interpretation of English puns. In: SemEval (2017)
9. Potash, P., Romanov, A., Rumshisky, A.: SemEval-2017 Task 6: #HashtagWars: Learning a sense of humor. In: SemEval. pp. 49–57 (2017)
10. Rajadesingan, A., Zafarani, R., Liu, H.: Sarcasm detection on Twitter: A behavioral modeling approach. In: Proc. of WSDM. pp. 97–106 (2015)
11. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in Twitter. Language resources and evaluation **47**(1), 239–268 (2013)
12. Shahaf, D., Horvitz, E., Mankoff, R.: Inside jokes: Identifying humorous cartoon captions. In: Proc. of KDD. pp. 1065–1074 (2015)
13. Yang, D., Lavie, A., Dyer, C., Hovy, E.: Humor recognition and humor anchor extraction. In: Proc. of EMNLP. pp. 2367–2376 (2015)
14. Zhang, R., Liu, N.: Recognizing humor on Twitter. In: CIKM. pp. 889–898 (2014)

[8] `http://eranik.me/humor-detection`