СЕНТИМЕНТНЫЙ АНАЛИЗ ТВИТОВ НА ОСНОВЕ СИНТАКСИЧЕСКИХ СВЯЗЕЙ

Адаскина Ю. В. (adaskina@gmail.com),
Паничева П. В. (ppolin86@gmail.com),
Попов А. М. (hedgeonline@gmail.com)

InfoQubes, Москва, Санкт-Петербург, Россия

Ключевые слова: сентиментный анализ, синтаксические связи, статистические алгоритмы, классификация текстов

SYNTAX-BASED SENTIMENT ANALYSIS OF TWEETS IN RUSSIAN

Adaskina Yu. V. (adaskina@gmail.com),
Panicheva P. V. (ppolin86@gmail.com),
Popov A. M. (hedgeonline@gmail.com)

InfoQubes, Moscow, Saint Petersburg, Russia

The paper describes our approach to the task of sentiment analysis of tweets within SentiRuEval—an open evaluation of sentiment analysis systems for the Russian language. We took part in the task of object-oriented sentiment analysis of Russian tweets concerning two types of organizations: banks and telecommunications companies. On both datasets, the participants were required to perform a three-way classification of tweets: positive, negative or neutral.

We used various statistical methods as basis for our machine learning algorithms and checked which features would provide the best results. Syntactic relations proved to be a crucial feature to any statistical method evaluated, but SVM-based classification performed better than the others. Normalized words are another important feature for the algorithm.

The evaluation revealed that our method proved to be rather successful: we scored the first in three out of four evaluation measures.

Key words: Sentiment analysis, syntactical relations, statistical methods, text classification

Introduction

In spite of being quite well explored by researches and businesses alike sentiment analysis remains to this day one of the most in-demand NLP tasks. Sentiment analysis had been applied on various levels, starting from the whole text level, then going towards the sentence level. Lately most of work has been focused on object-oriented and aspect based sentiment analysis, which is based on the assumption that different opinions can be expressed within one sentence. Today's research dwells not only on the development of automatic sentiment analysis algorithms, but also on evaluation methods. A number of independent bodies conduct evaluations, one of them being Dialogue Evaluation which is held in coordination with Dialogue—the international conference on computational linguisics. This is their third event devoted to sentiment analysis; the results of the first two are discussed in (Chetviorkin, Braslavskiy, Loukachevitch 2012) and (Chetviorkin, Loukachevitch 2013). This year's tasks was automatic evaluation of sentiment towards specific objects or their properties in different datasets (Loukachevitch et al. 2015).

This paper describes our approach to the task. We participated in the object-oriented sentiment analysis of Russian tweets concerning two types of organizations: banks and telecommunications companies. On both datasets, the participants were required to perform a three-way classification of tweets: positive, negative or neutral.

We applied SVM classification (Pedregosa et al. 2011) in our final experiments, although our preliminary results suggested that there was no significant difference between SVM and Naïve Bayes in this task. We used normalized words (further called lemmas) combined with syntactic relations as features. The latter are defined as triplets: source word, target word, relation type. Syntactic relations turned out to be crucial for any statistical method we used in our preliminary tests. All the methods we used showed better results on tweets about telecommunications companies, than on tweets about banks. The evaluation revealed that our method proved to be rather successful: we scored the first in three out of four evaluation measures.

Related work

(Pang, Lee, Vaithyanathan 2002) is generally considered the principal work on using machine learning methods of text classification for sentiment analysis; it explores the use of Naïve Bayes, Maximum Entropy and Support Vector Machines methods. The problem is further discussed in (Go, Bhayani, Huang 2009; Barbosa, Feng 2010 and Jiang et al. 2011), among others. Numerous research was dedicated to developing the ultimate feature set for each specific task to get the best result of automatic classification. Most common features are:

- word forms;
- normalized words;
- phrases;
- frequencies;
- TF-IDF;

- n-gram;
- binary occurrences;
- syntactic relations.

Syntactic information is less common than other parameters because clearly it presupposes a complicated and time-consuming stage of syntactic analysis. However, those experiments that involved dependency relations showed that syntax contributes significantly to both Recall and Precision of most algorithms. For the task of text classification in general see (Furnkranz, Mitchell, Rilof 1998), (Caropreso, Matwin, Sebastiani 2001), (Nastase, Shirabad, Caropreso 2006). (Matsuko et al. 2005) deal with a task very close to ours, sentiment classification based on syntactic relations. They parsed frequent sub-trees using two different algorithms, which is a more general approach than ours since we only used 'binary sub-trees', i.e. a pair of words in syntactic relationship. Another distinction is that we combined syntactic information with normal forms as features for machine learning based sentiment classification. (Bethard, Martin 2007) as well as (Zhang et al. 2007) used syntactic relations for the task of semantic relations mining. In (Zhao, Grishman 2005) the authors tackle the task of automatic context extraction, and syntactic relations are a key to their impressive 70% F-measure result.

The sentiment analysis of Twitter today is a full-fledged subtask within sentiment analysis per se. Due to the limited character count the analysis of tweets is closer to sentence-level sentiment analysis than the other blogging platforms. A number of papers discuss the specifics of Twitter sentiment analysis, see for example (Pak, Paroubek 2010; Kouloumpis, Wilson, Moore 2011; Jansen et al. 2009; Tumasjan et al. 2010).

Dataset and task description

We took part in a testing procedure of sentiment analysis systems with our algorithm. Full evaluation details are outlined in (Loukachevitch et al. 2015). The dataset consisted of training and evaluation sets, 10,000 tweets each. Both sets were divided into two subsets: 5,000 tweets about banks and 5,000 tweets about telecommunications companies. The training set had been manually annotated by SentiRuEval experts. This annotation included three-way annotation (negative, positive and neutral) for every company (seven telecommunications companies and eight banks) that was mentioned in the tweet. The test set had been annotated with neutrals for every company that was mentioned in the tweet. Within our task we needed to perform automatic sentiment analysis on the test set, which is either to retain a neutral annotation for the appropriate brand, or to change it to negative annotation or to a positive one. The evaluation set had been annotated by three assessors, and tweets where there was no agreement between the experts (at least two of the three), were excluded from the evaluation set. The total size of the evaluation set was 4,549 tweets for banks and 3,845 tweets for telecommunications companies.

Algorithm

We used InfoQubes morphosyntactic analyzer applied also in (Adaskina, Panicheva, Popov 2014). This is a commercial platform designed by our company. Its lemmatization module is based on Zaliznyak's Grammar Dictionary (Zaliznyak 1980); its syntactic module is a finite state machine, which parses word sequences and produces syntactic trees. An elaborated rule system (featuring 515 syntactic rules) is applied as input context-free grammar for the parser. Every syntactic rule joins two words or phrases into one higher-order phrase and sets respective syntactic relations. Thus, a constituency grammar is applied which in turn yields a dependency structure following a small number of rules. Only binary relations are allowed; each syntactic relation is characterized by three elements: source word, target word and relation type. In total, the system features 19 syntactic relations, their frequencies for both training datasets are presented in Table 1. In our parametrical model the relation (Argument) which has four subtypes (Subject, DirectObject, IndirectObject, PassiveSubject) is split into four different relations.

Table 1. Syntactic relation extracted for the training datasets

Relation Name	Occurrences in Telecom dataset	Occurrences in Banks dataset
Argument:DirectObject	2,778	2,372
Argument:IndirectObject	5,748	3,585
Argument:PassiveSubject	291	232
Argument:Subject	3,148	1,805
Attribute	6,814	6,682
Auxiliary	578	208
Circumstance	3,033	1,211
Coordinate	1,008	1,698
Determiner	687	239
Genitive	3,963	3,355
Identity	2,200	4,937
Infinitive	772	465
Modifier	707	294
Phrasal	1,519	959
Possessive	368	126
Preposition	6,582	4,554
Quantifier	501	605
Subordinate	226	77
Undefined	1,050	1,159

We tested simple word lemmas (unigrams), word lemma bigrams and syntactic relations as features for SVM and Naïve Bayes (Pedregosa et al. 2011) three-way classification (neutral, positive, negative) algorithms. In every experiment we normalized word forms according to the lemmatization module of the morphosyntactic tool. One of our underlying goals was to test the performance of syntax-based features in the sentiment analysis task. As optional settings we applied a negation marker provided by our morphosyntactic system. Negation marker in our system is a feature that marks cases where a negation particle is connected to the word. We also optionally removed from the parameter list everything that contained words denoting brands in question, implying that an overall brand bias could affect the result negatively. The features and their optional settings are summarized in Table 2.

Table 2. Feature descriptions

Features type	Feature text	Feature type	Options	Example	Comments
1	ВАРИАНТ	Lemma	No negation marker	Lemma *ВАРИАНТ*	Just normalized words
2	ВАРИАНТ|Argument|НЕТ|PassiveSubject	Syntactic relation	No negation marker	Passive subject relation *ВАРИАНТА НЕТ*	Syntactic relation of a certain type between two words. Relation 'Argument' also has four subtypes (Subject, DirectObject, IndirectObject, PassiveSubject), so the subtype is included
3	ВАРИАНТ|Attribute|ЭТОТ|	Syntactic relation	No negation marker	Attribute relation *ЭТОТ ВАРИАНТ,* words are not negated	Syntactic relation of a certain type between two words
4	КРУТОЙ|ВАРИАНТ	Bigram	No negation marker	Bigram *КРУТОЙ ВАРИАНТ*	Two adjacent words
5	ДРУГОЙ|ВАРИАНТ	Bigram	No negation marker	Bigram *ДРУГОЙ ВАРИАНТ*	Two adjacent words
6	ВАРИАНТ|0	Lemma	Negation marker included	Lemma *ВАРИАНТ,* not negated	A combination of normalized words and negation information; here the word is not negated

Features type	Feature text	Feature type	Options	Example	Comments
7	ВАРИАНТ|1	Lemma	Negation marker included	Lemma *ВАРИАНТ*, negated	A combination of normalized words and negation information; here the word is negated
8	ВАРИАНТ|1|Argument|НЕТ|0|PassiveSubject	Syntactic relation	Negation marker included	Passive subject relation *ВАРИАНТА НЕТ*, *ВАРИАНТ* is negated	A combination of syntactic relation and negation information; here one of words is negated
9	ВАРИАНТ|0|Attribute|ЭТОТ|0|	Syntactic relation	Negation marker included	Attribute relation *ЭТОТ ВАРИАНТ*, words are not negated	A combination of syntactic relation and negation information; here neither word is negated
10	КРУТОЙ|0|ВАРИАНТ|0	Bigram	Negation marker included	Bigram *КРУТОЙ ВАРИАНТ*, words are not negated	A combination of bigrams and negation information; here neither word is negated
11	ДРУГОЙ|0|ВАРИАНТ|1	Bigram	Negation marker included	Bigram *ДРУГОЙ ВАРИАНТ*, *ВАРИАНТ* is negated	A combination of bigrams and negation information; here one of words is negated

Preliminary results

We conducted some preliminary experiments applying ten-fold cross-validation to the training dataset only. Our text analysis algorithm consisted of sentiment classification described above and a rule-based algorithm of relevant brand identification. For every document we compiled a list of triplets: document id, brand id, sentiment score. We evaluated the results by computing the overall Precision, Recall and F1-measure over the lists of triplets obtained by text analysis and from the annotated information. Thus we also evaluated the relevant brand identification algorithm and included neutral class performance comparing to the SentiRuEval evaluation scheme. Results we obtained are presented in the following tables, and the highest scores are marked in bold; Table 3 refers to Telecom companies data, Table 4 to Banks data.

Table 3. Preliminary results for Telecom companies data, SVM

Features type	Experiment options		Evaluation		
	Negation marker	Brand name removal	Precision	Recall	F1-measure
Lemmas	–	–	0.7464	0.7482	0.7473
	+	–	0.7549	0.7567	0.7558
	–	+	0.7554	0.7571	0.7563
	+	+	0.7608	0.7625	0.7616
Relations	–	–	0.7275	0.5567	0.6308
	+	–	0.7228	0.5532	0.6267
	–	+	0.7196	0.5470	0.6216
	+	+	0.7215	0.5484	0.6231
Lemmas + relations	–	–	**0.7715**	**0.7734**	**0.7725**
	+	–	0.7692	0.7710	0.7701
	–	+	0.7675	0.7692	0.7684
	+	+	0.7632	0.7648	0.7640
Lemmas + relations, chi-square selection of 5000 best parameters	–	–	0.5865	0.5879	0.5872
Bigrams	–	–	0.7242	0.7077	0.7158
Bigrams + relations	–	–	0.7204	0.7220	0.7212
Bigrams + lemmas	–	–	0.7650	0.7668	0.7659
Bigrams + lemmas + relations	–	–	0.7684	0.7702	0.7693

Table 4. Preliminary results for Banks data, SVM

Features type	Experiment options		Evaluation		
	Negation marker	Brand name removal	Precision	Recall	F1-measure
Lemmas	–	–	0.9046	0.9061	0.9053
	+	–	0.9021	0.9036	0.9029
	–	+	0.9073	0.9087	0.9080
	+	+	0.9032	0.9046	0.9039
Relations	–	–	0.9040	0.8184	0.8591
	+	–	0.9080	0.8220	0.8628
	–	+	0.9040	0.8171	0.8583
	+	+	0.9066	0.8194	0.8608
Lemmas + relations	–	–	0.9059	0.9074	0.9066
	+	–	0.9047	0.9062	0.9055
	–	+	0.9083	0.9097	0.9090
	+	+	**0.9095**	**0.9108**	**0.9101**

	Experiment options		Evaluation		
Features type	**Negation marker**	**Brand name removal**	**Precision**	**Recall**	**F1-measure**
Bigrams	–	–	0.8968	0.8949	0.8959
Bigrams + relations	–	–	0.8957	0.8971	0.8964
Bigrams + lemmas	–	–	0.9021	0.9036	0.9029
Bigrams + lemmas + relations	–	–	0.9026	0.9041	0.9033
Lemmas + relations, chi-square selection of 5000 best parameters	–	–	0.8257	0.8269	0.8263

Our preliminary experiments have shown that a combination of lemmas and syntax relations yield the best results for both datasets, while negation and brand name removal options do not considerably affect the performance. That result is consistent with our initial hypothesis that syntactic features should improve the performance. Bigrams and lemmas are almost as good as relations and lemmas. Naïve Bayes classification has confirmed these tendencies with a small decrease in performance. We also tried excluding some features, but the results were unsatisfactory. The tables above include scores for feature selection of 5,000 best parameters, and one can see that this decreased the resulting score rather significantly. Apart from that, we tried tf-idf value, but it also reduced our evaluation metrics. It appears that the data might be too sparse for the weighting factors to work: they probably would have been useful for an experiment with a larger training set where the frequency of each parameter would be higher, and there would be fewer parameters with unique values.

SentiRuEval testing results

For the final experiment within the testing procedure framework we have chosen SVM classification with lemmas and syntactic relations as features, we have also removed brand names from the feature set as an option. We have also performed an out of competition evaluation of the lemmas-based algorithm. Table 5 below represents evaluation results, the numbers in the last column refer to our experiment types ('lemmas', 'lemmas+relations') or the results by other participants (indicated by their number). In italics is our result obtained out of competition. As the main quality measures the evaluation team used two variations of F-measure: F-micro and F-macro, for details see (Loukachevitch et al. 2015). The best result in each category is marked in bold, and, as one can see from the data, our method scored the first in three out of four evaluation measures.

Table 5. Final evaluation results

Domain	Measure	Baseline	Participant results	Participant identifier
Telecom	Macro F	0.182	**0.488**	**lemmas+rels**
			0.483	lemmas+rels, brands removed
			0.480	3
			…	…
			0.469	*lemmas*
			0.465	*lemmas, brands removed*
	Micro F	0.337	**0.536**	**lemmas+rels**
			0.536	lemmas+rels, brands removed
			0.528	10
			…	
			0.512	*lemmas*
			0.514	*lemmas, brands removed*
Banks	Macro F	0.127	**0.360**	**4**
			0.352	10
			0.345	*lemmas*
			0.345	*lemmas, brands removed*
			0.343	lemmas+rels, brands removed
	Micro F	0.238	**0.366**	**lemmas+rels, brands removed**
			0.364	lemmas+rels
			0.363	*lemmas*
			0.362	*lemmas, brands removed*
			0.343	8

There is a notable difference in performance between the preliminary experiments and the testing procedure results, which is naturally justified by a difference in evaluation methods: we have applied F-measure to all the documents in the former case, while in the latter the neutral documents were excluded..

These results are only partially consistent with our preliminary results and our initial hypothesis: on the Telecom dataset the performance of lemmas and relations combined outdoes lemmas only by approx. 2 per cent in micro and in macro F-measures. On the Banks dataset the result is inconclusive: micro F-measure is better by about 0.3 per cent than lemmas and relations combined, but macro F-measure is about 0.2 per cent better with lemmas only. The Banks dataset is also characterized by overall lower performance when the neutral class is not accounted for in the evaluation, contrary to our preliminary experiments yielding higher performance with 'Banks' comparing to 'Telecom'. This fact and the inconsistency of the 'Banks' results distribution (almost the same performance for lemmas and lemmas with relations) suggest that the algorithms applied can't achieve reliable performance with the modest volumes of negative- and positive-class data.

The closest best results in the SentiRuEval scheme were obtained with techniques involving rule-based fact-extraction, MaxEnt and SVM classifiers over various feature sets mostly including word and letter n-grams.

Conclusions

We have applied a syntax-based statistical algorithm to sentiment analysis tasks in two different topics yielding very high performance results comparing to other techniques. We have used straightforward classification features, slightly improving the performance of a simple lemma approach with syntactic relations or not affecting it where the sparsity of data wouldn't allow for reliable high results: the issue that needs to be further addressed. We have used an elaborate morphosyntactic parser, which had proven useful for another semantic task (Adaskina, Panicheva, Popov 2014).

With sparse and modest-sized data SVM appears to be the best classification method; negation or brand-name semantics do not affect the performance much, though we believe that syntactic relations would convey most of the information carried by the negation option. It also appears that the sparsity of data does not allow for effective feature filtering, which could be an option if we boost feature occurrence by, for example, substituting words with semantic classes.

References

1. *Adaskina Yu. V., Panicheva P. V., Popov A. M.* (2014), Semi-Automatic Lexicon Augmenting Based on Syntactic Relations [Poluavtomaticheskoye popolneniye slovarey na osnove sintaksicheskikh svyazey], Proceedings of Internet and Modern Society Conference, Saint Petersburg, pp. 271–276.
2. *Barbosa L., Feng J.* (2010), Robust Sentiment Detection on Twitter from Biased and Noisy Data, Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, pp. 36–44.
3. *Bethard S., Martin J. H.* (2007), CU-TMP: Temporal Relation Classification Using Syntactic and Semantic Features, Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, pp. 245–248.
4. *Caropreso M. F., Matwin S., Sebastiani F. A.* (2006), Learner-Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization. Amita G. Chin (ed.), Text Databases and Document Management: Theory and Practice, Idea Group Publishing, pp. 78—102.
5. *Chetviorkin I., Braslavskiy P., Loukachevich N.* (2012), Sentiment Analysis Track at ROMIP 2011, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2012"], Bekasovo, pp. 1–14.
6. *Chetviorkin I., Loukachevitch N.* (2013), Evaluating Sentiment Analysis Systems in Russian, Proceedings of BSNLP workshop, ACL, Prague, pp. 12–17.

7. *Furnkranz J., Mitchell T. M., Rilof E.* (1998), A Case Study in Using Linguistic Phrases for Text Categorization on The WWW, Proceedings of the AAAI Workshop on Learning for Text Categorization, Madison, US, pp. 5–12.
8. *Go A., Bhayani R., Huang L.* (2009), Twitter Sentiment Classification Using Distant Supervision, Technical report, Stanford.
9. *Jansen, B. J., Zhang, M., Sobel, K., Chowdury, A.* (2009), Twitter power: Tweets as electronic word of mouth, Journal of the American Society for Information Science and Technology, 60(11), pp. 2169–2188.
10. *Jiang L., Yu M., Zhou M., Liu X., Zhao T.* (2011), Target-dependent Twitter Sentiment Classification, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, US, pp. 151–160.
11. *Kouloumpis E., Wilson, T., Moore J.* (2011), Twitter sentiment analysis: The good the bad and the omg! Artificial Intellingence, pp. 538–541.
12. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Ju., Ivanov V., Tutubalina H.* (2015), Sentirueval: Testing Object-Oriented Sentiment Analysis Systems In Russian.
13. *Matsumoto S., Takamura H., Okumura M.* (2005), Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees. Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 301–311. Springer, Heidelberg.
14. *Nastase V., Shirabad J. S., Caropreso M. F.* (2006), Using Dependency Relations for Text Classification. In Proceedings of the 19th Canadian Conference on Artificial Intelligence, Quebec City, pp. 12–25.
15. *Pak A., Paroubek P.* (2010), Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC, Valetta, pp. 75–100.
16. *Pang B., Lee L., Vaithyanathan S.* (2002), Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 2002, Vol. 10, Stroudsburg, US, pp. 79–86.
17. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.* (2011), Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 12(Oct), pp. 2825–2830.
18. *Tumasjan, A., Sprenger, T. O., Sandner, P., Welpe, I.* (2010), Predicting elections with twitter: What 140 characters reveal about political sentiment. Proceedings of ICWSM, Washington, US, pp. 178–185.
19. *Zaliznjak, A. A.* (1980) Grammatical dictionary for Russian language. Rus. jaz, Moscow
20. *Zhang M., G. Zhou, A. Aw* (2008), Exploring Syntactic Structured Features Over Parse Trees for Relation Extraction Using Kernel Methods, Information Processing and Management, vol. 44, issue 2, pp. 687–701.
21. *Zhao S., Grishman R.* (2005), Extracting Relations with Integrated Information Using Kernel Methods, Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor, US, pp. 419–426.