Methods for Automatic Sentiment Detection

Anatoly A. Gurin [0000-0003-3751-7604]

Plekhanov Russian University of Economics,
36 Stremyanny lane, Moscow, 115998, Russia
anatoly196674@gmail.com

Abstract. Semantic analysis has great potential applications in various fields of science and the national economy**.** Much of the information in the world is not structured, so there is the problem of processing and extracting useful data. Natural language processing is a very complex process. On November 23, 2017, as part of the 2017 Runet Prize award ceremony, the HOT-LIST 2018 was presented, that identified the main digital trends of 2018 and presented a list of trendsetters in 10 technology, innovation and business areas.There were included 10 leading companies in 10 areas and the first area is AI technologies. AI technology has become a key technology trend in 2018, and the volume of global investment in these technologies and products based on them in excess of $ 1 billion.[25]

More than 180 private companies working on projects in the field of AI technologies have been purchased during the period 2011-2018. According to IDC Customer Insights & Analysis.[25] During the period 2011-2018 it was purchased more than 180 private companies working on projects of AI technologies. According to forecasts of Frost & Sullivan, by 2022 artificial intelligence market will grow to $ 10 billion using machine learning technologies and natural language recognition in advertising, retail, finance and health.[25]

Keywords: sentiment analysis, rule-based approach, sentiment lexicon, machine learning, dictionary approaches for determining the sentiment of text, comparison of methods for automatic sentiment determination.

1 Introduction

Semantic analysis has great potential applications in various fields of science and the national economy. Currently, it is used in the monitoring, analysis, signal systems, document management systems, advertising platforms, and much more. Tonality analysis of the text is to extract from the text of opinions and emotions, as well as their subsequent processing, refers to methods of content analysis and is a means of exploring subjectivity in natural language.[1]

Much of the information in the world is not structured, so there is the problem of processing and extracting useful data. Natural language processing is a very complex process. There exist numerous methods. This may be the use of lexical and grammatical structures, along with an estimated lexicon. Besides, machine-learning techniques can be used to solve such problems. In this approach to solving the problem of senti-

Proceedings of the 10th International Scientific and Practical Conference named after A. I. Kitov "Information Technologies and Mathematical Methods in Economics and Management (IT&MM-2020)", October 15-16, 2020, Moscow, Russia

ment analysis, necessary set of texts as a training sample. Machine learning with the teacher needs a set of pre-marked text reviews.

On November 23, 2017, as part of the 2017 Runet Prize award ceremony, the HOT-LIST 2018 was presented, that identified the main digital trends of 2018 and presented a list of trendsetters in 10 technology, innovation and business areas.There were included 10 leading companies in 10 areas and the first area is AI technologies.

AI technology has become a key technology trend in 2018, and the volume of global investment in these technologies and products based on them in excess of $ 1 billion.[25]

More than 180 private companies working on projects in the field of AI technologies have been purchased during the period 2011-2018. According to IDC Customer Insights & Analysis.[25]

During the period 2011-2018 it was purchased more than 180 private companies working on projects of AI technologies. According to forecasts of Frost & Sullivan, by 2022 artificial intelligence market will grow to $ 10 billion using machine learning technologies and natural language recognition in advertising, retail, finance and health.[25]

Dynamics of artificial intelligence is based on five fundamental technologies:

- machine learning, [6]
- in-depth training, [2]
- computer vision, [2]
- natural language processing, [6]
- machine reasoning and strong artificial intelligence. [4]

The main drivers of the market will be the sectors of consumer products, business services, advertising and defense. Processing market natural language (NLP) and products on its basis is estimated by experts in the area of $ 8 billion in 2018 and will grow to $ 40 billion by 2025. [25] The main drivers will be the increasing demand for more advanced level of user experience, increased use of smart device, the growth of investments in health care, the growing use of network and cloud-based business applications and the growth of M2M-technology. NLP market growth is constrained by factors such as the presence of the gap in perception / understanding / recognition of textual information between man and machine, the shortage of personnel and training programs for researchers in the field of NLP, as well as the complexity of machining and understanding of the context and meaning of the text. [3] It is also one of the challenges in the segment of natural language processing is the creation of universal language models and architectures that will address a variety of work tasks with the text with a single system. That is a system that will "understand" text information and be able to communicate with the person as it would make the other person who has read the text and having some amount of knowledge[11]. Certain restrictions are applied directly to the understanding of the Russian language. In this case, the quality of understanding depends on many factors: language, national culture of the interlocutor, etc. One of the major technology trends in the segment of natural language processing for today - is to use machine-learning techniques to reduce labor costs for text layout, machine learning methods without a teacher or partial involvement of teachers, active methods of machine learning, etc. [8] High efficiency in dealing with language processing tasks shown as vector representation of words and other language constructions - that is, deep machine learning and neural networks. Therefore, many natural language processing tasks today are solved with the use of vector representations and deep learning of neural networks. Also, one of the trends the last time - is to use the algorithm of knowledge transfer (Transfer Learning), in which the NLP-trained models to solve simple tasks with the use of large amounts of data. Further, these pretraining models are used for other, more specific tasks. [5]

2 The main approaches for determining tonality

Tonality analysis is generally defined as one of the problems of computational linguistics, i.e. meant that we could find and classify the tone using natural language processing tools (such as a tagger, parsers, etc..). Make a big generalization, it is possible to divide the existing approaches into the following categories [18]:

1. Approaches based on rules;
2. Approaches based on dictionaries;
3. With the teacher machine learning;
4. Machine learning without a teacher.

The first type of system consists of a set of rules applying to the system concludes the tone of the text. Many commercial systems using this approach, even though it is costly, because for the good work it is necessary to make many the rules of the system. [7] Often the rules are tied to a specific domain (e.g. "theme restaurant") and the change of the domain ("review of the camera") is required to re-make the rules. [15] However, this approach is the most accurate in the presence of a good rule base, but it is not interesting for the study. Approaches based on dictionaries, use so-called tonal dictionaries for text analysis. In the simplest form, tonal vocabulary is a list of words with the value of the tone for each word. [24] To analyze the text, one can use the following algorithm: first, every word in the text to assign it a value of tonality from the dictionary (if it is present in the dictionary), and then calculate the overall tone of the text. Calculate the overall tone of a variety of ways. The simplest of them - the average of all values. A more complex - to train a classifier (e.g. a neural network.). [17]

Machine learning with a teacher It is the most common method used in the research. Its essence is to teach the machine classifier on the collection of pre-marked-up text, and then use the resulting model for the analysis of new documents. It is about this method. [23]

Machine learning without a teacher It is probably the most interesting and at the same time the least accurate method of analysis pitch. One example of this method may be automatic clustering documents. [23]

Machine learning with a teacher. The process of creating the tone analysis system is very similar to the process of creating other systems using machine learning [9]

- It needs to assemble a collection of documents for training the classifier;
- Each document from the training necessary to present the collection in the form of a feature vector;

- For each document, needs to specify the «correct answer», tone type (e.g. positive or negative), and for those responses to be trained classifier;
- Classification algorithm selection and training of the classifier;
- Use the resulting model;

The number of classes that are shared key is usually set of system specifications. For example, the customer is required for the system to distinguish three kinds of tone: «positive», «neutral», «negative». The studies usually consider the problem of binary classification key, i.e., only two classes: the «positive» and «negative».[10] Classification tone for more than two classes - this is a very difficult task. Even with the three classes is very difficult to achieve good accuracy regardless of the approach used. The most interesting method is a method based on dictionaries. [19]

2.1 Approaches based on dictionaries. SentiStrenght software

According to research published in the article Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology [26]. It was tested various algorithms for determining the strength of positive mood for 1041 comments with an extended set of functions and 10-fold cross-validation (in descending order of indicators of the strength of positive mood). [20] Except for SentiStrength, the results are the average of 4 runs of different random test / training sections and for an optimal number of features. The results are shown in Table 1.

Algoritm	**Characters**	**Accuracy**	**Accuracy +/- 1 class**	**Corr.**	**Abs mean %**
SentiStrength (standard configuration, 30 runs)	-	60.6%	96.9%	.599	22.0%
Simple logistic regression	700	**58.5%**	96.1%	**.557**	**23.2%**
SVM (SMO)	800	**57.6%**	**95.4%**	**.538**	**24.4%**
J48 classification tree	700	**55.2%**	**95.9%**	.548	24.7%
JRip rule-based classifier	700	**54.3%**	96.4%	**.476**	**28.2%**
SVM regression (SMO)	100	**54.1%**	97.3%	**.469**	**28.2%**
AdaBoost	100	**53.3%**	**97.5%**	**.464**	**28.5%**
Decision table	200	**53.3%**	96.7%	**.431**	**28.2%**
Multilayer Perceptron	100	**50.0%**	*94.1%*	**.422**	**30.2%**
Naive Bayes	100	**49.1%**	**91.4%**	**.567**	**27.5%**
Baseline	-	**47.3%**	**94.0%**	**-**	**31.2%**
Random	-	**19.8%**	**56.9%**	**.016**	**82.5%**

In terms of the power of negative emotions, most methods give very similar results, and some give better results than SentiStrength. Although the SentiStrength accuracy is 72.8%, this is only 2.9% better than the baseline, some other methods have similar accuracy levels, and SVM is significantly more accurate. SentiStrength is the most accurate of the methods when one class error is allowed and has the highest correlation with human coding results. In theory, none of the methods should be worse than the baseline, but this can be due to the optimization of the training set, not the estimate set. Overall, it might seem that SentiStrength is not very good at recognizing negative emotions, but this is a difficult task for the short texts analyzed here. Also, the average percent absolute error for the random category is over 100% due to the predominance of "1" as the correct category for negative sentiment. In this way systems based on the vocabulary approach are no worse than systems using machine learning, and sometimes even better, it depends on the specific task.

2.2 Machine learning with a teacher. Common rules

This is the most common method. Its essence is to train a classification algorithm based on a collection of documents. The classes of which are known in advance. [22] Advantages:

- High accuracy in determining the tonality;
- The problem of dependence on a specific subject area is solved by training the classifier based on a sample from this area, since the classifier itself selects features that affect the sentiment, many studies are carried out in order to improve accuracy.

Disadvantages:

- A marked-up collection of texts is needed (markup is a very time-consuming process).

The algorithm of this approach can be described as follows:

- First of all, it needs to choose a collection of documents based on which the classifier will be trained;
- Each document must be presented as a vector of features (aspects);
- Further, each document must be assigned the correct type of sentiment;
- It is necessary to choose a classification algorithm and method for training the classifier;
- Application of the resulting model.

It is necessary to decide how many classes and what type of classification will be used. It is difficult to get high results when using flat classification. Research shows that the best results are obtained using hierarchical classification. All documents from the training set must be n-dimensional vectors of aspects. [21] The quality of the results directly depends on which set of characteristics will be used. The most common ways of presenting documents are bag-of-words or n-grams form. [13]

There are two classification algorithms naive Bayes classifier and support vector machine (SVM). After choosing the classification algorithm and training the classifier, the results are assessed with cross-validation process help. [12]
Formula for finding accuracy:

$$P = \frac{correctly\ extracted\ opinions}{total\ number\ of\ opinions\ found\ by\ system} \quad (1)$$

In the case of cross-validation, the data is split into k parts, then the model is trained on k-1 parts of the data, and the rest is used for testing.

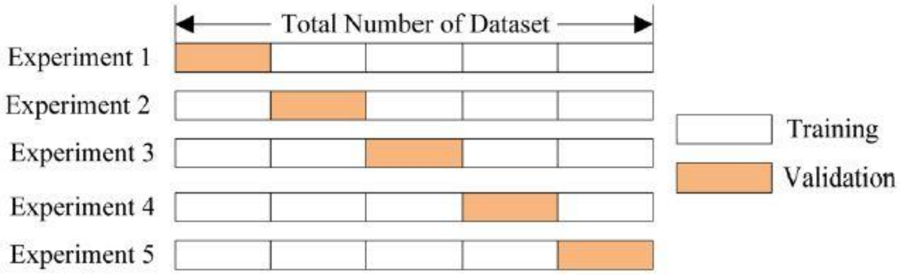

Fig. 1. Cross-validation process

3 Review of existing systems, determining the tone of the text

IBM Watson Explorer is a set of models which allow one to look up information in a text, to select entities and relationships.

Baidu ERNIE 2.0 - a framework for understanding natural language is available in English and Chinese languages, supports the inference definition of semantic similarity, recognition of named entities, the key of the analysis and comparison of the questions and answers.

Apple platform for natural language processing - a framework for language identification, tokenization, lemmatization, parts of speech tagging and identification of named objects.

Facebook bAbI - a platform for the automatic interpretation of texts, as well as a set of datasets to test algorithms for natural language understanding.

Facebook FastText - framework for the classification of text, highlight key words and named entities.

Tencent NLP - an open platform with features semantic analysis that provides an API for the development of NLP-systems solutions and applications of natural language processing.

AliReader - technology analysis of unstructured text, intelligent search and retrieve information from a variety of documents used in many products Alibaba.

Russian companies are leading the development in the field of NLP, on the market in several categories. First, it is the search engines and the company, which for many years engaged in text technology «Yandex», ABBYY, Mail.ru, PROMT and RCO (part of the group Rambler).

The second category - large corporations, which only in the last 3−4 years began to form their competence in the field of AI. For example, Sberbank, «Tinkoff Bank» MTS. All of them have achieved impressive results, even though do basically the technology for internal use.

"Yandeks.Toloka"- crowdsourcing platform for the collection and processing of data for the ML-projects, training of search algorithms and neural networks, the development of speech technologies and computer vision. The "Cleanup" there are more than 5 million 20 thousand, and performers. Customers. Collected assessment are used to develop voice assistants and chat bots and research in different domains.

"Sberbank"- monitoring and automatic content analysis of news about the 1000 partner banks in Russian. NLP-decision ABBYY selects a meaningful message, categorizes news on various risk factors and collects relevant data dossiers of the banks.

Just AI Conversational Platform- enterprise-level platform for the development of conversational chat-bots and assistants who understand natural language. Chat bots started in the platform, to solve complex business challenges: customer support, recruitment and training of staff, ordering and selling of goods.

PROMT Analyzer SDK- a component for information analysis systems. Allows you to automatically analyze Big Data in different languages, highlights the fact, mentioned persons, organizations, events, and other entities, to determine the tone of the statements and documents.

EUREKA ENGINE (BRAND ANALYTICS)- High system of linguistic analysis module type text that allows to extract new knowledge and facts from unstructured data in huge volumes of real-time.

RCO Fact Extractor SDK- a tool of the computer analysis of textual information. The package is designed for developers of information-analytical and search engines.

RCO Text Categorization Engine- Developer library for information retrieval systems, allowing based on lexical profiles define the text belonging to a given set of categories to get the number of entries and the position of the selected term in the text.

Project iPavlov- overcoming technological barriers in the field of meaningful human-machine communication in natural language through the creation and introduction to business practice tools, reducing the threshold of entry into the market of text dialogue systems. The goal is realized through the following tasks: research and development of network architectures for working with text in a natural language.

Creating views analysis system is a challenging task, but doable if there is data for training and pre-defined theme. [14] When using machine learning is important to test different options to pick the ones that work best on the test data. It needs to test different algorithms for classification (NB, SVM), a set of features (unigrams, bigrams, the character N-gram), signs the weighing function. There are many ways to improve the classification key, such as the use of tonal dictionaries, additional linguistic features (eg, parts of speech), and general methods for the improvement of machine learning (boosting, Bagging and others.). [16]

4 Results

We have decided to compare two methods. It is approaches based on dictionaries and machine learning with a teacher. For machine learning we applied an algorithm:

- First of all, it needs to choose a collection of documents based on which the classifier will be trained;
- Each document must be presented as a vector of features (aspects);
- Further, each document must be assigned the correct type of sentiment;
- It is necessary to choose a classification algorithm and method for training the classifier;
- Application of the resulting model.

The participles of working program are:

- Lemmatization of a document - bringing all words in a document to their initial form using the pymorphy2 morphological analyzer, removing punctuation marks, service parts of speech, words that contain letters of the Latin alphabet.
- Allocation of text features - the document is compared with 3 numbers, which are a numerical characteristic of its emotional coloring, calculated using the TF-IDF formula. Weights are calculated for using unigrams (document words one at a time), bigrams (a combination of two words) and trigrams (a combination of 3 words).
- Determining the document class ("positive", "negative") using the Naive Bayesian classifier.

For approaches based on dictionaries we have followed algorithm:

- Data cleaning. All text is scanned and extra characters are removed;
- All words are reduced to their initial form, using OpenCorpora;
- After using tonal dictionaries as well as dictionaries expressions and idioms, words are weighted;
- In addition to calculating negative, neutral and positive tonalities, emotions are calculated using OCC model;
- After determining the key and calculating the values 16
- emotions, the program evaluates the tonality of the text. The calculation of emotions is one of the important processes, because their identification improves the operation of the sentiment determination algorithm, and at the same time checks it for errors.

A dataset was selected for testing our programs. There were 30000 tweets that had been determined. Program based on machine learning with a teacher made a mistake 32% of the time. Program based on dictionaries approaches made a mistake 25% of the time. In this way approach based on dictionaries works better than approach based on machine learning with a teacher.

5 Discussion and further directions

Sentiment analysis is a field of computational linguistics devoted to the automatic identification of assessments, emotions of a person regarding entities in the text, or the identification of a general assessment of the emotionality of a statement.

Tonality is the emotional color expressed in the text. This analysis is used both for commercial purposes and for solving scientific research problems.

There are two types of opinions - simple and comparative. Most of the works devoted to this area of research are engaged in identifying simple opinions, since comparative ones are extremely difficult to analyze.

Most often, there are 6 tasks of analyzing the emotional coloring of the text - the extraction of objects, aspects, the author and their classification; extraction of time and its standardization; determination of tonality; opinion conclusion.

The main problems that arise when analyzing sentiment include the dependence of sentiment on the subject area, the use of emotive vocabulary in neutral sentences, sarcasm, the dependence of sentiment on the user who reads the message, as well as the expression of sentiment without using emotionally colored words. These problems can be eliminated with varying degrees of success in the process of sentiment analysis.

There are four main approaches to the analysis of the sentiment of texts:

- method using rules;
- method using a dictionary of emotional vocabulary;
- method based on teaching with a teacher,
- method based on unsupervised learning.

Each of the approaches has advantages and disadvantages.

The supervised learning method was discussed in detail, as it is one of the most popular approaches to sentiment analysis. This approach has five main steps. At the first stage, a training sample is prepared. Each document is represented as a vector of features (most often it is either a "bag of words" or n-grams) with further assignment of weights to each element of the vector. Then a classification algorithm is selected. In this chapter, we looked at how the two most efficient algorithms work the naive Bayesian classifier and the support vector machine. To evaluate the performance of the model, either completeness and accuracy can be calculated or cross-validated.

Method using a dictionary of emotional vocabulary is also popular. For example SentiStrength software. It was developed by Mike Telwall, Kevan Buckley and their colleagues at the University of Wolverhampton in 2010. The program evaluates the sentiment strength of short messages simultaneously on two scales (positive and negative) from 1 to 5 and from -1 to -5. This system is based on the use of emotional vocabulary and corrective rules. The program was developed based on messages from the MySpace social network. The dictionary was partially supplemented with vocabulary from the LIWC dictionary. The chapter describes in detail the process of creation and the algorithm of SentiStrength.

The algorithm of this tool is more effective than other methods in dealing with the analysis of the positive sentiment of short informal texts. The result of the program was evaluated using the accuracy and the correlation coefficient between the expert estimates and the program estimates. Later, the developers made several attempts to improve the program for negative sentiment by expanding the original data.

6 Conclusion

Thus, we can conclude that all methods of determining the tonality are good. At the moment, there is no universal solution and everything depends on the task and needs of the customer. For example, the rule-based approach gives excellent results, but it is limited in application, as well as in the selected domain (topic). A sharp change of domain strongly affects the definition of the sentiment, and it does not work very well when there is an intersection of completely different topics. Machine learning approaches are also not perfect, but they are popular due to the technological order and development information society. There are a lot of adds in these approaches that, if applied correctly, it gives good results. Dictionary approaches also have their drawbacks and show good results in some tasks, and not very well in others. It is happened due to vocabulary approaches are not universal, but these approaches, combined with others, such as machine learning, may well provide a very accurate definition of sentiment, regardless of the subject matter and data structure.

References

1. Asur Sitaram and Bernardo A. Huberman. Predicting the future with social media. Arxiv preprint arXiv: 1003.5699, 2010.
2. Babbar Rohit, Partalas Ioannis, Gaussier Eric, Amini Massih-Reza. On Flat versus Hierarchical Classification in Large-Scale Taxonomies.
3. Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of the Seventh conference on International Language Resources and Evaluation, pp. 2200-2204.
4. boyd, d. (2008). Taken out of context: American teen sociality in networked publics. University of California, Berkeley, Berkeley.
5. boyd, d. (2008). Why youth (heart) social network sites: The role of networked publics in teenage social life. In D. Buckingham (Ed.), Youth, identity, and digital media, pp. 119-142. Cambridge, MA: MIT Press.
6. Bradley, M. M., & Lang, P. J. (1999). Affective Norms for English Words (ANEW): Stimuli, instruction manual, and affective ratings (Tech. Report C-1). Gainesville: University of Florida, Center for Research in Psychophysiology.
7. Brill, E. (1992). A simple rule-based part of speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing, pp. 152-155.
8. Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.P. Measuring User Influence in Twitter: The Million Follower Fallacy. Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM), Washington, May 2010.
9. Choi, Y., & Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 793-801.

10. Das, S., & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), Bangkok, Thailand, July 22-25, last accessed 20.02.2020: http://sentiment.technicalanalysis.org.uk/DaCh.pdf.
11. Derks, D., Bos, A. E. R., & von Grumbkow, J. (2008). Emoticons and online message interpretation. Social Science Computer Review, 26(3), pp. 379-388.
12. Fox, E. (2008). Emotion science. Basingstoke: Palgrave Macmillan, p. 127.
13. Freitas A.A., de Carvalho A.C.P.L.F. (2007) Research and trends in data mining technologies and applications: tutorial on hierarchical classification with applications in bioinformatics.
14. Fullwood, C., & Martino, O. I. (2007). Emoticons and impression formation. The Visual in Popular Culture, 19(7), pp. 4-14.
15. Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text (IDA 2005). Lecture Notes in Computer Science, 3646, pp. 121-132.
16. Ghazi Diman, Inkpen Diana, Szpakowicz Stan. Hierarchical versus Flat Classification of Emotions in Text. Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 140-146, Los Angeles, California, June 2010.
17. Joshi Mahesh, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. Movie reviews and revenues: An experiment in text regression. In Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference (NAACL 2010), 2010.
18. Jurafsky Daniel, Martin James H. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition. Pearson Education International, 2009. 1024 pp.
19. Kan D. Rule-based approach to sentiment analysis. Sentiment Analysis Track at ROMIP, 2011.
20. Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage.
21. Kukich, K. (1992). Techniques for automatically correcting words in text. ACM computing surveys, 24(4), pp. 377-439.
22. Liu Bing. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, May 2012.
23. Liu Bing. Sentiment Analysis Tutorial. AAAI-2011, San Francisco, USA.
24. Liu Yang, Huang Xiangji, An Aijun, Yu Xiaohui: ARSA: a sentiment-aware model for predicting sales performance using blogs. SIGIR 2007: pp. 607-614.
25. Research by Frost & Sullivan in the interests of Up Great Technology Competitions (organized by RVC, ASI and the Skolkovo Foundation).
26. Sentiment Strength Detection in Short Informal Text. Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai. Statistical Cybermetrics Research Group, School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK.