Your Paper has been Accepted, Rejected, or whatever: Automatic Generation of Scientific Paper Reviews

Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao

Department of Engineering and Architecture, University of Trieste, Trieste, Italy

Abstract. Peer review is widely viewed as an essential step for ensuring scientific quality of a work and is a cornerstone of scholarly publishing. On the other hand, the actors involved in the publishing process are often driven by incentives which may, and increasingly do, undermine the quality of published work, especially in the presence of unethical conduits. In this work we investigate the feasibility of a tool capable of generating fake reviews for a given scientific paper automatically. While a tool of this kind cannot possibly deceive any rigorous editorial procedure, it could nevertheless find a role in several questionable scenarios and magnify the scale of scholarly frauds.
A key feature of our tool is that it is built upon a small knowledge base, which is very important in our context due to the difficulty of finding large amounts of scientific reviews. We experimentally assessed our method 16 human subjects. We presented to these subjects a mix of genuine and machine generated reviews and we measured the ability of our proposal to actually deceive subjects judgment. The results highlight the ability of our method to produce reviews that often look credible and may subvert the decision.

1 Introduction

Peer review, i.e., the process of subjecting a work to the scrutiny of experts in order to determine whether the work deserves publication, is a keystone in scholarly publishing. The review process should ensure that a published paper is of high scientific quality, which in its turn preserves the reputation of the corresponding publishing venue and improves the prestige of its author. On the other hand, peer review is just a piece of broader process involving several entities whose incentives may or may not actually drive the overall process toward those ideal goals. Authors are increasingly subject to strong pressures in the form of research evaluation procedures in which the indicators that play a key role are often mostly numerical [1]. Reviewers tend to be overworked and often receive little credit for their hard work [2], while at the same time being interested in increasing some counter of program committees or editorial boards in which they are involved. Commercial publishers may find in scholarly publishing excellent opportunities for profit [3], even in the form of journals with little or no scrutiny:

a periodically updated list of *predatory publishers* has grown by 50 times in the last 5 years, including 923 publishers in its latest release [4].

While there is no doubt that most published research follows a rigorous and honest path, it is evident that actors involved in research may now find ways to maximize their personal benefits disregarding the ideal objective of the scientific environment as a whole, by following practices that are questionable or simply fraudulent [5, 6]. Unfortunately, this claim is not a mere theoretical possibility. Questionable operators have emerged that run bogus journals and conferences which have no other purpose than generating profit while uttering worthless scientific literature [7]. Supposedly peer-reviewed journals accept for publication papers that have been randomly generated [8] or publish papers which clearly have not been proof-read by anyone [9]. Misbehaving researchers attempt to inflate their records by ghostwriting papers on nonexistent research [10]. Not surprisingly, the critical reviewing step has been exploited as well. Computer intrusions on the editorial system of a major commercial publisher have forced the publisher to retract several published papers [11]. In the last few years, hundreds of published papers have been retracted by several commercial publishers in many independent events [12, 13, 14], due to the discovery of reviews fabricated by the authors themselves which provided journals with suggested reviewers along with fake contact information which actually routed communication to the authors or their colleagues.

In this work, we investigate the feasibility of more fraud opportunities in the form of a procedure for *automatic generation of fake reviews*. We propose a method for generating automatically text which (a) looks like the typical scientific paper review, (b) is tailored to the specific paper being reviewed, and (c) conveys a recommendation specified as input. A tool that is capable of generating fake reviews systematically and at *no cost* may be misused in several ways. Busy people which want to be involved in as many reviewing committees as possible might choose a recommendation and then generate reviews very quickly, perhaps without even reading the paper or after just a superficial look. Predatory publishers might attempt to improve their credibility by sending many reviews to authors. Of course, reviews generated by our tool will certainly be detected as being fake by any decent editorial process. On the other hand, as pointed out above, perverse incentives and unethical conducts might find a role for a tool of this kind, which may potentially magnify the scale of frauds in the reviewing process in several ways. In this respect, it is important to keep in mind that a few years ago Springer and IEEE retracted more than 120 published papers which were computer-generated nonsense [15]. Our proposed tool could find more constructive applications, though. For example, the steering committee of a conference could inject fake reviews in the discussion phase without informing the program committee and then observe the outcome.

Our proposed method constructs a review tailored to a specific paper, with a specified recommendation, based solely on the paper text and a corpus of reviews written by humans for other papers. A key aspect of our proposal is that it builds upon a relatively small knowledge base (some tens of reviews)

while commonly used methods for text generation, such as Artificial Neural Networks (ANN), typically require a very large amount of data in order to build an effective generative model. Applying those methods in the context of scientific review generation is difficult because of the difficulty in finding a large amount of samples of scientific reviews, in particular, of negative reviews.

An important contribution of our work is the experimental campaign performed involving human subjects. We performed an *intrinsic* evaluation aimed at assessing the ability of our method to generate reviews which look like as being written by a real human reviewer. Moreover, we performed an *extrinsic* evaluation aimed at assessing the impact on the decision about accepting or rejecting a paper under review. Although our experimental campaign is not a replica of a real editorial process and thus may provide only a preliminary assessment, our results do provide interesting insights.

2 Related work

To the best of our knowledge, no method for the automatic generation of reviews of scientific papers has been proposed before. From a broader point of view, our proposal is a form of Natural Language Generation (NLG), which is widely used in many different fields such as spoken dialogue systems [16], machine translation [17], and as a mean for creating editorial content by turning structured data into prose [18].

A notable use of NLG for scientific purpose, which is particularly relevant to our work, is the software SCIgen[1]. This tool generates pdf files consisting of syntactically correct random text which is formatted like a scientific publication, including randomly generated figures, plots, and code fragments. Later and independently from its creators, SCIgen has been used in order to test the submissions standard of conferences and to prove that nonsense papers may actually be published, even by respected publishers [15]. This phenomenon has been investigated also in [19], which studies the spread of fakes and duplicates through notable publishers. The fact that a tool which was born as a "toy" for Computer Science researchers led to actual malicious behaviors suggests that other types of cheating may arise, including the creation of false reviews: this consideration is indeed the main motivation of our work.

Our work proposes a *corpus-based* NLG method. Corpus-based methods aim at training text generation rules automatically from text examples of the desired text generator output. An example of corpus-based method applied to text generation in dialogue is the work in [20]. The cited work proposes a class-based n-gram language model (LM) that improves over template-based and rule-based text generation systems. Belz [21] proposes a corpus-based probabilistic generation methodology and apply it to the automatic generation of weather forecast texts. The work in [22] assesses a new model for NLG in dialogue systems by maximizing the expected reward using reinforcement learning.

[1] `http://pdos.csail.mit.edu/scigen/`

A different approach to NLG is based on Artificial Neural Networks (ANN). Kukich [23] implemented a stock reporter system where text generation is done at phrase level using an ANN-based approach. A recent work demonstrated the effectiveness of Recurrent Neural Networks (RNN) for natural language generation at character level [24]. A variant of RNN, Long Short-Term Memory (LSTM) [25], proved its ability to generate characters sequences with long-range structure [26]. The authors of [27] showed the ability of a LSTM framework to automatically generate rap lyrics tailored to the style of a given rapper. Zhang and Lapta [28] proposed an RNN-based work for generating Chinese poetry. Beyond unbounded text generation, LSTM for NLG has also been used in the generation of image descriptions [29, 30, 31] and in the generation of descriptive captions for video sequences [32].

All the generative methods based on neural networks require a huge amount of learning data, usually orders of magnitude more than the amount of data that we could find in our scenario (i.e., scientific reviews). Methods for *data augmentation* capable of decreasing the amount of learning data required for training a neural network effectively certainly deserve investigation in our context [33].

3 Our approach

The problem consists in generating, given a paper a and an overall recommendation $o \in \{\text{accept}, \text{neutral}, \text{reject}\}$, a review r which (i) appears as generated by a human (ii) for the paper a and (iii) which expresses a recommendation o for a. In our work, we assume that the paper a is a plain text which consists of the concatenation of the paper title, abstract and main content.

Our method requires a set R of real paper reviews, i.e., each review $r \in R$ has been written by humans. We pre-process each review in R as follows: (i) we split the document in a sequence $\{t_1, t_2, \dots\}$ of tokens according to the Penn-Treebank procedure; (ii) we execute a *Named-entity Recognition* (NER)[2] [34] on the token sequence; and (iii) we execute a *Part-of-Speech* (POS) annotation[3] [35] on the token sequence; finally (iv) we classify each token in $\{t_1, t_2, \dots\}$ as being or not being a specific term, according to an heuristic procedure (see below).

When generating a review for a paper a with a specified recommendation o, our method performs 3 steps, described below in full detail: (i) it builds a set S of sentences from reviews in R and replaces each specific term in each sentence with a specific term of a; (ii) it removes from S the sentences which express a sentiment which is not consistent with o; (iii) it reorders and concatenates the sentences in S obtaining a review for a.

Specific terms identification. With this procedure, we aim at identify the *specific terms* of a document d—i.e., those terms which are relevant to d. To this end, we defined a simple heuristic. Let $\{t_1, t_2, \dots\}$ the sequence of tokens for d, where each token has been annotated with NER and POS taggers. A token

[2] `http://nlp.stanford.edu/software/CRF-NER.shtml`
[3] `http://nlp.stanford.edu/software/tagger.shtml`

$t \in \{t_1, t_2, \dots\}$ is a specific term if it meets all the following criteria: (i) t has been annotated as a noun (`NN`) or as an adjective (`JJ`); (ii) the length in characters of t is at least 2; (iii) t contains at least one letter.

Specific terms replacement. In this step, we aim at constructing a set S of review sentences tailored to a. To this end, we proceed as follows, starting with $S = \emptyset$. For each review $r \in R$, we split the review in a set S_r of sentences. We obtain (according to the procedure described above) the set W'_a of specific terms of a, retrieve the set W'_r of specific terms of r, and set $W_a = W'_a \setminus W'_r$ and $W_r = W'_r \setminus W'_a$. Then, for each sentence $s_r \in S_r$, we generate a random mapping from items in the set W^s_r of specific terms of W_r which occur in s_r to items in W_a such that: (a) each item in W^s_r is mapped to exactly one item in W_a, (b) no items in W^s_r exist such that they are mapped to the same item in W_a, and (c) for each item w^s_r mapped to an item w_a, the POS and NER annotations of w^s_r are the same of respective annotations of w_a. If such mapping is possible, we replace each occurrence of a term of W^s_r in s_r with the mapped term in W_a and add the modified sentence to S; otherwise, we proceed to the next sentence.

In other words, after this procedure, S contains all the suitable sentences generated by iterating the term replacement procedure for all the reviews in R.

Sentiment analysis. In this step, we aim at selecting the sentences of S which express a sentiment consistent with the specified overall recommendation o. To this end, we apply a pre-trained Naive Bayes sentiment classifier[4] [36] to each sentence $s \in S$, basing on the assumption that a positive sentiment can be associated with an accept recommendation, a negative sentiment with a reject recommendation, and a neutral sentiment with a neutral recommendation.

After the application of the sentiment classifier, we retain in S only the sentences for which the outcome is consistent with o.

Sentences reordering. In this step, we aim at generating the final output of our method (the automatically generated review) by selecting, reordering, and concatenating a subset of sentences of S. The rationale for the selection and reordering is to obtain a review (a) whose length is realistic, w.r.t. a typical review, and (b) which has an overall structure which resembles a typical review—e.g., an opening sentence, some considerations, a conclusive remark.

Concerning the reordering, we based on the assumption that sentences may be classified as suitable for opening part, central content, and closing part. Accordingly, we built a classifier which takes as input a single sentence and outputs a label in $\{\text{opening}, \text{central}, \text{closing}\}$. We took the general purpose text classifier based on maximum entropy[5] described in [37] and trained it using all the sentences of the reviews in R, which we automatically labeled as follows: if the sentence was the first sentence in its review, we associated it with the label opening; otherwise, if it was the last sentence, we associated it with closing; otherwise, we associated it with central.

[4] `http://sentiment.vivekn.com`
[5] `http://nlp.stanford.edu/software/classifier.shtml`

When generating a review, we apply the classifier to each sentence in S and then randomly select 1 opening sentence, 3 central sentences, and 1 closing sentence. Finally, we concatenate those 5 sentences and obtain the review for a.

4 Experimental evaluation

We performed two experimental evaluations involving human subjects for assessing our proposed method ability to generate reviews which (a) look like as they have been written by real human reviewers for the specified paper, and (b) can affect the decision about accepting or rejecting the specified paper. That is, we performed an intrinsic evaluation and an extrinsic evaluation, respectively.

We built a dataset composed of 48 papers and 168 reviews, which we obtained from the F1000Research, Elifescience, Openreview and PeerJ web sites—which publish reviews of accepted papers along with corresponding full texts—and from our lab publication records; we used the reviews of the dataset as the set R while running our method. Moreover, for the purpose of performing our evaluations, we associated an overall recommendation (i.e., a label in $\{\text{accept}, \text{neutral}, \text{reject}\}$ with each review in the dataset. Since the sources we considered vary in the way, if any, they classify reviews according to overall recommendation, we proceeded as follows. If a review was explicitly associated with an overall recommendation by its author, we associated it with the suitable label—e.g., positive recommendations to accept, negative recommendations to reject, and all the other recommendations to neutral. Otherwise, if a review was not explicitly associated with an overall recommendation, we considered the outcome of the publishing process which, for published papers, was always acceptance.

In order to provide a comparison baseline for our review generation method, we designed and built a simple baseline generation method based on Markov chains. To this end, we trained a second order Markov chain, operating on tokens, on all the reviews in the dataset: before the training, we added a special token t_{end} at the end of each review. When generating a review with the baseline method, the specified paper a and the overall recommendation o are not considered and the following steps are performed. First, a review in the dataset is randomly chosen and its first two tokens are fed into the Markov chain generative model. Then, the generative model is run until the token t_{end} is obtained. Finally, the output is obtained by concatenating all the generated tokens.

In our experimentation, we involved a number of human subjects, who were asked to examine the generated reviews and then to answer some questions. In order to gain more insights about our method effectiveness, we grouped the subjects according to their presumed familiarity with scholarly publishing, resulting in 3 classes. The experienced class is composed of professors, PhD student, and postdocs; the intermediate class is composed of undergraduate students; the novice class is composed of all the remaining subjects (who were anyway sufficiently proficient with English).

4.1 Intrinsic evaluation

In the intrinsic evaluation, we built a number of forms, each showing the title of a paper a randomly chosen from our dataset and a set of 10 reviews randomly sampled for the following sets: (a) the real reviews in the dataset actually related to a, (b) the real reviews in the dataset not related to a, (c) a set of reviews generated using the baseline method, and (d) a set of reviews generated using our method with a and a random overall recommendation o as input. Since the size in characters of the real reviews can widely vary, we limited the number of sentences presented to the subject to 5, as for our generated reviews, randomly sampled from the corresponding reviews while maintaining the original ordering. We asked the subject to say, for each review in the form, if "it appeared as a genuine review written by a human reviewer for the paper with the shown title". We gathered results from 16 subjects—5 novice, 3 intermediate, and 8 experienced.

Figure 1 shows the key findings of the intrinsic evaluation: the figure plots the percentage of positive answers (on the y axis) to the form questions for each kind of review (bar group) and for each class of subjects (bar fill pattern). It can be seen that our method generates reviews that are considered as written by a human in almost one case on three—the figure being greater for novice subjects an smaller for experienced subjects. Moreover, the deceiving ability is larger than the baseline: approximately 30% vs. 10%. Concerning the real reviews, Figure 1 shows that, as expected, they are properly recognized $\approx 85\%$ of the times: this finding suggests that the truncation of real reviews does not severely affect their appearance.

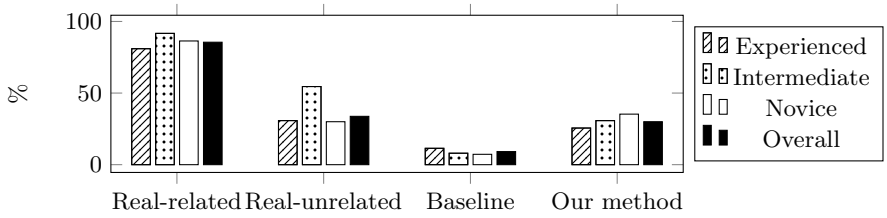

Fig. 1. Percentage of reviews considered as written by a human for the specified paper.

4.2 Extrinsic evaluation

In the extrinsic evaluation, we built a number of forms, each showing the title of a paper a randomly chosen from our dataset and a set of 3 reviews randomly sampled for the sets described at points a, b, and d in the previous section. Real reviews were possibly limited in length as in the intrinsic evaluation. The form also showed, next to each review, the corresponding overall recommendation. We asked the subject to answer the following two questions: 1. "basing on these 3 reviews, would you recommend to accept or reject the paper?"; 2. "while taking

your decision, in which order the 3 reviews influenced you?" We gathered results from 13 subjects—3 novice, 3 intermediate, and 7 experienced.

Table 1 summarizes the key findings of the extrinsic evaluation. In the left portion the table shows, for each subject class and for all the subjects, the number of forms in which at least a real and a generated reviews were discordant w.r.t. the recommendation (Discordant column), the number of discordant forms for which the subject took a decision in line with the generated reviews (and hence against the real reviews, Subverted column), and the ratio among Subverted and Discordant. In the right portion it shows the number of forms, for each kind of reviews, in which a review of the corresponding type were stated to be the most influencing by the subject; moreover it shows the percentage of forms in which the generated reviews were stated to be the most influencing.

Subject class	Subverted	Discordant	%	Our method	Original	Others	%
Experienced	4	16	25.0	10	21	4	28.6
Intermediate	4	15	26.7	11	18	14	25.6
Novice	5	21	23.8	11	25	9	24.4
Overall	13	52	25.0	32	64	27	26.0

Table 1. Results of the extrinsic evaluation (see text).

The most interesting, and somewhat surprising, finding is that in the 25% of cases the decision of an experienced subject agreed with the generated reviews and disagreed with the real reviews: from another point of view, through a generated review we were able to manipulate the outcome of the (simulated) peer review process. Table 1 also shows that, in 26% of cases, a generated review was stated to be the most influencing by the subjects.

5 Conclusions

We proposed a method for the automatic generation of scientific reviews. The method is able to generate a review of a given research paper with a specified overall recommendation. To this end, it performs multiple steps aimed at generating reviews which resemble human written reviews and hence might potentially induce the reader to accept or reject the reviewed paper.

A key contribution of our work is the experimental evaluation, which involved 16 human subjects. The results show that in $\approx 30\%$ of cases a generated review is considered genuine by the human subjects; moreover, in about 1 among 4 cases, we were able to manipulate the outcome of a (simulated) peer review process through generated reviews which we mixed with genuine reviews.

Beyond these promising results, our proposal needs further investigation and, in this respect, we plan to compare it with other NLG methods, such as ANN, for which, however, a much larger amount of data need to be collected. Finally, it could be interesting to investigate if and how an ontology can improve the review generation process.

References

1. Bartoli, A., Medvet, E.: Bibliometric evaluation of researchers in the internet age. The Information Society **30**(5) (2014) 349–354
2. Csiszar, A.: Peer review: Troubled from the start. Nature **532**(7599) (apr 2016) 306–308
3. HEFC: Identification and dissemination of lessons learned by institutions participating in the research excellence framework (ref) bibliometrics pilot. Technical report, Higher Education Funding Council for England (2009)
4. Beall, J.: List of predatory publishers 2016. `https://scholarlyoa.com/2016/01/05/bealls-list-of-predatory-publishers-2016` Accessed: 2016-29-04.
5. Bowman, J.D.: Predatory publishing, questionable peer review, and fraudulent conferences. American journal of pharmaceutical education **78**(10) (2014)
6. Dadkhah, M., Alharbi, A.M., Al-Khresheh, M.H., Sutikno, T., Maliszewski, T., Jazi, M.D., Shamshirband, S.: Affiliation oriented journals: Don't worry about peer review if you have good affiliation. International Journal of Electrical and Computer Engineering **5**(4) (2015) 621
7. Butler, D., et al.: The dark side of publishing. Nature **495**(7442) (2013) 433–435
8. Eldredge, N.: Mathgen paper accepted! Technical report, That's Mathematics (2012)
9. Oremus, W.: This is what happens when no one proofreads an academic paper. `http://www.slate.com/blogs/future_tense/2014/11/11/_crappy_gabor_paper_overly_honest_citation_slips_into_peer_reviewed_journal.html` (2016)
10. Qiu, J., Schrope, M., Jones, N., Borrell, B., Tollefson, J., Kaplan, M., Lovett, R.A., Dalton, R., Merali, Z.: News publish or perish in china. Nature **463** (2010) 142–143
11. Reller, T.: Faking peer reviews. Technical report, Elsevier Connect (2012)
12. Fischman, J.: Fake peer reviews, the latest form of scientific fraud, fool journals. Technical report, The Chronicle of Higher Education (2012)
13. Ferguson, C., Marcus, A., Oransky, I.: Publishing: The peer-review scam. Nature **515**(7528) (nov 2014) 480–482
14. Callaway, E.: Faked peer reviews prompt 64 retractions. Nature (aug 2015)
15. Noorden, R.V.: Publishers withdraw more than 120 gibberish papers. Nature (feb 2014)
16. Wen, T.H., Gasic, M., Mrkšić, N., Su, P.H., Vandyke, D., Young, S.: Semantically conditioned lstm-based natural language generation for spoken dialogue systems. (September 2015) 1711–1721
17. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. (2014) 3104–3112
18. Wright, A.: Algorithmic authors. Communications of the ACM **58**(11) (2015) 12–14
19. Labbé, C., Labbé, D.: Duplicate and fake publications in the scientific literature: how many scigen papers in computer science? Scientometrics **94**(1) (2013) 379–396
20. Oh, A.H., Rudnicky, A.I.: Stochastic natural language generation for spoken dialog systems. Computer Speech & Language **16**(3) (2002) 387–407
21. Belz, A.: Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. Natural Language Engineering **14**(04) (2008) 431–455

22. Rieser, V., Lemon, O.: Natural language generation as planning under uncertainty for spoken dialogue systems. In: Empirical methods in natural language generation. Springer (2010) 105–120
23. Kukich, K.: Where do phrases come from: Some preliminary experiments in connectionist phrase generation. In: Natural language generation. Springer (1987) 405–421
24. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: INTERSPEECH. Volume 2. (2010) 3
25. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780
26. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
27. Potash, P., Romanov, A., Rumshisky, A.: Ghostwriter: Using an lstm for automatic rap lyric generation. (2015) 1919–1924
28. Zhang, X., Lapata, M.: Chinese poetry generation with recurrent neural networks. In: EMNLP. (2014) 670–680
29. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3128–3137
30. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632 (2014)
31. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3156–3164
32. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014)
33. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)
34. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics (2005) 363–370
35. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, Association for Computational Linguistics (2003) 173–180
36. Narayanan, V., Arora, I., Bhatia, A.: Fast and accurate sentiment classification using an enhanced naive bayes model. In: Intelligent Data Engineering and Automated Learning–IDEAL 2013. Springer (2013) 194–201
37. Manning, C., Klein, D.: Optimization, maxent models, and conditional estimation without magic. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials-Volume 5, Association for Computational Linguistics (2003) 8–8