Classification of Bots and Gender using Topic Unigrams

Astrid Fleig, Lisa Geyersbach, Melissa Göhler, Patricia Kurz, Paul Limburg, Dirk Labudde and Michael Spranger
University of Applied Sciences Mittweida
Mittweida, Germany
Email: spranger@hs-mittweida.de

***Abstract*—In social networks such as Twitter, author profiling plays a big role. It is especially interesting to differentiate between accounts from humans and bots and to make a prediction about the age and the gender of human users. The information can be helpful to analyze possible manipulations, networks and crimes. This paper presents an approach to differentiate between bots and humans, as well as the gender for the human accounts using Tweets. For each sub-problem, a linear Support-Vector Machine (SVM) was used and different feature and featuresets were tested. The analysis showed that the topic model is the best feature for all categories. For this feature, the term frequencies of the most important terms of the topics were used. In comparison to other approaches, this approach could increase the performance. More precisely, only with this feature it was possible to reach accuracies between 99.7% and 100%.**

Index Terms*— *Author Profiling; Bot Detection; Gender Detection; Twitter; Spanish; English.

I. Introduction

In social networks like Twitter, accounts run by bots are common [1] [2]. Some can be recognized at first sight, others stay undiscovered [3]. Depending on their use, bots can be a positive and helpful extension to the Twitter experience or can be harmful and deceptive [4]. Positive examples are weather bots that regularly post the weather, like @EmojiWeatherUSA, temperature in a certain area, like @EVER_WEATHER or tsunami warnings, like @NWS_PTWC. However, negative examples are bots that try to deceive people, use spam or harm users with malicious links [5] [6]. So called social bots use their platform to react to real peoples' tweets and spread commercial, political or ideological opinions. Using good deception methods, like being able to interact in conversations, those bots stay undetected by common Twitter users and can influence large groups of people. For example, they can bring users to believe in certain misinformation or vote for a specific politician in the next election [7]. By identifying bots automatically, a lot of those negative influences can be prevented.

Another interesting task is knowing the gender of Twitter users. This is being explored for several reasons. A Twitter user's gender can be used for forensic, criminological, political or phenomenological analysis [8]. For example, the amount of Tweets about an upcoming election can be analyzed regarding the author's gender or the members of a criminal network can be inspected further. Getting this information about the author's gender can be challenging because it does not need to be disclosed on the user profile. Thus, finding a way to accurately guess the gender of a user would be helpful. In this paper the focus is on the genders female and male.

The field of author profiling is well studied and even a competition regarding the described classification tasks was held with good results in 2019 by PAN [9].

In this paper, a new feature set is tested to identify whether a Twitter user is a bot or human and, afterwards, whether the human users are female or male. This feature set consists of a combination of extracted topics, bigrams and other surface-level features. Especially, the topics have not yet been used before. In addition to that the transferability of the system from one language to another is explored. Merely the content of each author's tweets will be used for these tasks.

In Section II, other approaches are discussed. Afterwards, in Section III an overview of the data used is given. In Section IV the different pre-processing steps are described, as well as the the feature extraction and the classification process. Finally, in Section V the results are presented and discussed, while in Section VI a brief conclusion is given and possible future work discussed.

II. Literature

The differentiation of bots and humans, as well as women and men, is an important problem, which is addressed often. There are many ways to approach this topic.

Different literature uses different classifiers for this task, e.g. Naive Bayes (NB) [2], Random Forest [2] [10] or logistic regression [11] [12] [13]. The best results are obtained by using a support-vector machine [4] [14] [15]. To solve the multi class problem a combination of multiple SVMs was used [4].

Using a combination of a Convolutional Neural Network (CNN) and a Recurrent Neural Networks (RNN) [16] is a different approach but obtains slightly worse results.

Popular features for those kinds of classification problems are word- and character n-grams, especially, word unigrams and word bigrams as well as character-3-to-5-grams [4] [16] [11] [12] [14] [15]. Partly, a TF-IDF-weighting with sublinear term frequency was used [14]. The latter has been proven to be useful especially for the gender differentiation.

Profile data like follower ratio and tweet frequency can be used for the differentiation if available [2] [10]. This data obtained good results for classifying authors in humans, bots and cyborgs [2]. It was also shown, that humans tweet more irregular and in undetermined time intervals, which results in entropy being a helpful feature for detecting humans. Tweet length was also already used for this task [10].

Furthermore, hashtags and user-mentions were proven to be relevant features [4]. Also, to detect typical bot behavior

 ISBN: 978-1-61208-891-4

especially hashtag- and user-mention count as well as the number of retweets and hyperlinks can be helpful [13] [17].

Literature shows a connection between bots and spam [2]. Spam can be understood as the lack of topic variance or extreme persistence of one topic. This can be shown by repeatedly tweeting the same tweet or merely posting tweets containing only one or few specific topics. Because most bots are focusing on specific topics and are recognizable by that behavior, topics can possibly be used as a feature. Finding and defining topics in tweets is an interesting subject, e.g. to filter tweets by factual relevant tweets. Approaches for this problem are topic detection using Latent Dirichlet Allocation (LDA) [18] or unigram clustering resulting in network-graphs with relations [19]. Using LDA can also show how intensely authors focus on one topic by using certain words frequently [17]. In connection to this, sentiment analysis can be used as well [17].

Differentiating between the two examined genders obtained the best results by using emoji lists, punctuation trigrams, Part-Of-Speech (POS)-trigrams, document sentiment or different wordlists as features [11]. In addition to that POS-sequence-patterns, the differentiation of writing styles and the consideration of word endings obtained good results for gender detection in texts [20].

Based on the described literature, in this paper linear SVMs and a feature set containing hashtag-, user-mention- and retweet count, document length, punctuation marks and word unigrams as well as bigrams is being used. Special attention is given to the topics of the tweets.

III. Data Set Description

The data used was originally provided for the author profiling task of the PAN competition in 2019. Overall, data sets for two languages, English and Spanish, were provided. Each of the data sets includes 100 tweets per author, as well as the ground truth. [9] The data was split into training and validation data as suggested by the PAN organizers [9].

The data sets are balanced in terms of their class distribution, as shown in Table I. Additionally, the original test data set was used to test the model developed in this work under the same conditions as in PAN 2019. The test data have the same characteristics and class distribution as the training data set.

IV. Methods

In this paper, an SVM based classification approach is chosen for both, the classification of bots and humans, as well as females and males. The approach is based on successful approaches discussed in the literature.

The overall procedure is shown in Figure 1 and consists of several consecutive and parallel sub-tasks, which are necessary for extracting the different feature sets.

Each task shall be explained in more detail below.

TABLE I
Overview of the class distribution of all data sets

	Spanish				English			
	b	h	f	m	b	h	f	m
train	1040	1040	520	520	1440	1440	720	720
	$\sum 2080$				$\sum 2880$			
val	460	460	230	230	620	620	310	310
	$\sum 920$				$\sum 1240$			
test	900	900	450	450	1320	1320	660	660
	$\sum 1800$				$\sum 2640$			

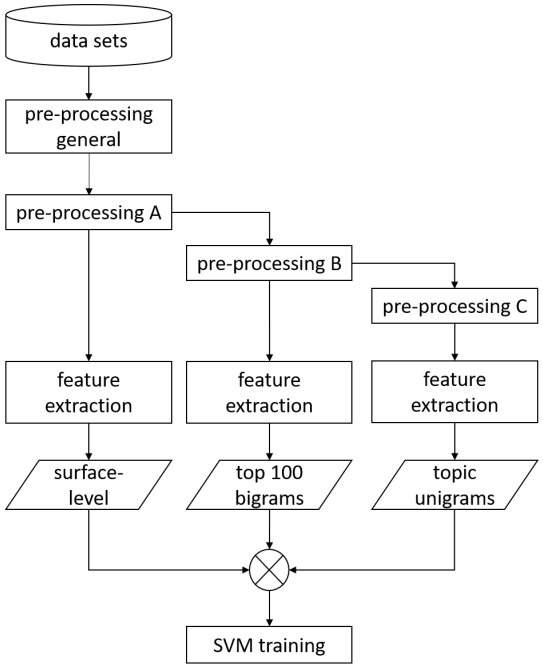

Fig. 1. General procedure for the classification tasks.

A. Pre-processing

The general pre-processing task is independent of the type of feature to be extracted. It consists of several steps, which were partly taken from [4]. In particular, every word was converted to lowercase and numbers, digits and isolated letters were erased. During the feature extraction process it was found that links are not useful for the classification, because there are no significant differences between the tweets of the different categories in regards to their number and content. Therefore, they were deleted. Furthermore, stop words for the respective languages were removed. However, for the Spanish data English stop words were removed as well because some English words or phrases are in the data. Finally, all spaces, resulting from the general pre-processing step were deleted.

For the extraction of certain feature sets special pre-processing was necessary. For the extraction of the bi-grams all special characters and punctuation marks, but hashtags

and user-mentions were deleted (pre-processing B). They were included because the relation between hashtag or user-mention and an additional word can be important, e.g. the term 'Trump' is often used with the hashtag '#politics'. The same pre-processing steps were used for the extraction of surface level features, however, exclamation marks, question marks and ellipses were not deleted (pre-processing A). Lastly, pre-processing step C deletes # and @ symbols. For the extraction of topic unigrams, pre-processing steps A and C were combined. Hashtags and user-mentions were deleted for the extraction of topic-unigrams as it is irrelevant whether a term is used inside or outside a hashtag or user-mention, since only the frequency of each term is counted.

B. Extracting Surface-Level Features

After pre-processing step B some surface-level features were extracted. First, the number of words of all tweets an author has written was analyzed. This proved to be useful in combination with other surface-level features.

Further, the number of retweets, user-mentions, and hashtags per author were considered as features, as well as the number of punctuation marks. During the feature extraction it became apparent that bots in this data set use more user-mentions than humans. In addition bots use twice as many hashtags. In the same way, retweets were helpful when differentiating between bots and humans. The opposite seems to be true for the gender classification. Here, these features show very similar usage in both gender categories.

As a stand-alone feature ellipses were not helpful for the tasks. However, in combination with other features the number of support vectors can be lowered by using them as a feature. Overall, humans use exclamation and question marks more often than bots. Furthermore, female and male authors are predominantly different in their use of ellipses.

C. Extracting Topical Terms and Bigrams

As it turned out, the most efficient feature can be created by using unigram topic models.The feature creation process is shown in Figure 2.

After pre-processing B, a Document-Term-Matrix (DTM) was created with a minimum Term Frequency (minTF) and minimum Document Frequency (minDF) of two. An LDA was executed on these DTMs. After running multiple tests and adjusting the topic count, with an amount of seven topics the best result were achieved on the given training data. For further steps and to prevent overfitting, only the top 20 words of each of the extracted topics were used to form the topic-unigram feature.

In order to obtain the frequencies of the extracted topical terms, a second DTM was created with a minTF of two and a minDF of ten. This DTM was used to count the occurrences of the extracted top 20 topical words for each author.

During the processing of the test and validation data similar DTMs were created without the restriction of minTF and minDF.

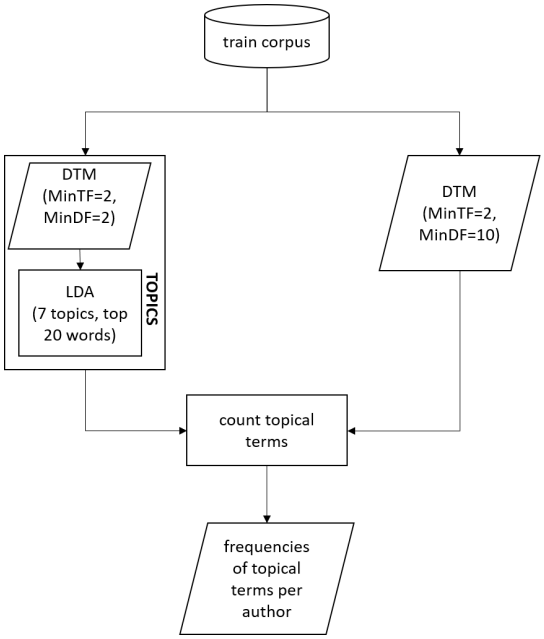

Fig. 2. Process of determining topical term frequency.

As shown in Figure 3 topics as features were very successful for differentiating both, bots and humans, as well as females and males. Bots turned out to be mostly talking about work, weather and news but also about advertising-related topics like gaming or YouTube. Human authors on the other hand were interested more in sports, politics, social networks, technology and free time activities.

Furthermore, it was observed, that female authors mostly wrote about topics like social networks and private events and male authors rather wrote about free time activities or politics.

The top 100 word bigrams were extracted from the data sets as a final feature. During this process, a document frequency minimum of 10 and a term frequency minimum of 2 was chosen to prevent overfitting [4] [11]. The extraction process was similar to the one of the topic-unigrams.

D. Feature-Evaluation

In order to evaluate the predictability of the different features an SVM was trained for each of them, using a ten-fold cross-validation after scaling the features. In Figure 3 the results of this evaluation are shown.

For both tasks the topic-unigrams and top 100 bigrams turned out to be the best features with accuracies of up to 100%. Generally, the single feature accuracies do not differ largely between the languages Spanish and English.

Retweet- and user mention count are, with accuracies of approximately 80%, also good features for differentiating between bots and humans. Ellipsis as a feature has the worst discrimination power for differentiating human and bots, yet it is the best surface-level feature for differentiating between females and males. Generally, surface-level feature have a slightly less discrimination power in the female/male differentiating task.

 ISBN: 978-1-61208-891-4

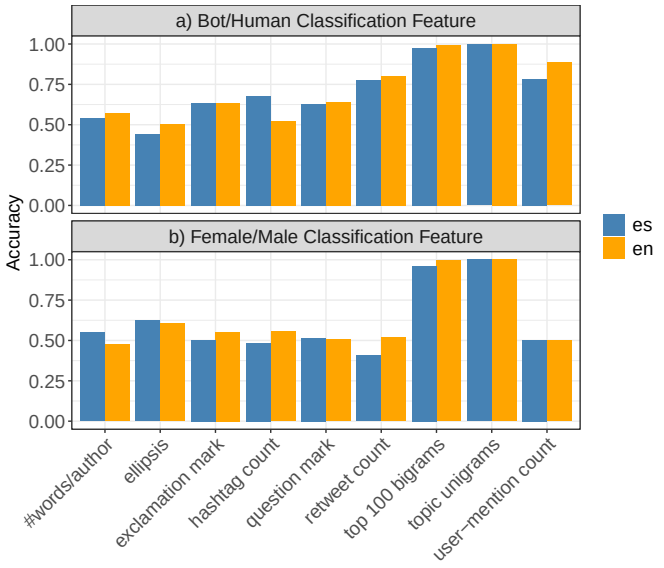

Fig. 3. Accuracy overview of the individual features of the Bot/Human classification in a) and of the Female/Male classification in b)

E. Experimental Design

In order to test the approach presented in this paper, four different feature sets were created. Firstly, a feature set that includes all of the discussed features (aF). Secondly, a feature set that only incorporates topic unigrams (U). Lastly, feature sets that are specific for each classification (Fk). These include for the Bot/Human classification a combination of document length, hashtag-, user-mention-, retweet-, and question mark count as well as topic-unigrams and top 100 bigrams and for the Female/Male classification a combination of document length, user-mention-, question mark-, exclamation mark- and ellipsis count as well as topic-unigrams and top 100 bigrams. For each of the classification tasks, a linear SVM was trained. Furthermore, for comparison, three baseline approaches were considered. For one baseline a Naive Bayes classifier (NB) was used and trained with the surface-level features hashtag-, user-mention count and document length. These surface-level features were chosen, since the accuracies on both, the Spanish and English data set, were very similar for each classification task (Figure 3). This baseline was used to set a minimal accuracy limit that definitely had to be surpassed and was not supposed to be especially challenging. It was utilized as an orientation what accuracy only few features can achieve. The other two baselines were taken from the literature.

[4] and [15] show the challenges that are supposed to be surpassed. The results in [4] serve as a baseline because the used approach is similar to the one used in this paper and, thus, is a good reference. Furthermore, [15] is the paper with the best accuracies for the given task. Thus, the goal in this paper was to surpass these accuracies.

V. Results and Discussion

A comparison of all three baselines and results of the classification with the given Test (TD) and Validation Data (VD) sets are presented in Table II. This Table shows results for all possible combinations of feature- and data sets. As explained in Subsection IV-E the feature sets are all of the discussed features (aF), only topic unigrams (U) and specific feature sets for the classification at hand (Fk). The table is also split into the languages Spanish and English, as well as the sub-problems Bot/Human (B\H) and Female/Male (F\M).

TABLE II
Final results in comparison to three baselines.

	Spanish		English	
	B\H	F\M	B\H	F\M
Baseline NB	0,713	0,5717	0,8677	0.5548
Baseline [4]	0,91	0,78	0,92	0,82
Baseline [15]	0,9333	0,8172	0,9360	0,8356
SVM+aF+TD	1	1	0,9992	1
SVM+Fk+TD	0,9978	1	1	1
SVM+U+TD	0,9967	1	1	1
SVM+aF+VD	0,985	0,9933	1	1
SVM+Fk+VD	0,9906	0,9922	0,9996	1
SVM+U+VD	0,9917	0,9922	0,9985	1

The most important result is that all baselines were surpassed by at least 7%. The difference between the accuracies reached with the naive bayes and all SVM results is especially great. This is a good result, since this baseline was supposed to be the lowest limit that had to be exceeded. Furthermore, the baselines by [4] and [15] were surpassed, too.

It can be noticed that the Female\Male differentiation of the English data is always at an accuracy of 100%. A reasonable cause for this outcome may be overfitting, even though precautions were taken to prevent this. Additionally the 100% accuracy of the Spanish TD concerning this task, may also be explained by overfitting.

Furthermore, there is only a minimal decline in the accuracies from the test to the validation data set. The results of the validation data set in comparison to the test data set dropped in no case more than 1.5%. This maximal loss in accuracy occurs in the Bot\Human differentiation of the Spanish data between the test and the validation data set using all features (aF). However, between the test and the validation data set the results even increased by 0.08%, in the case of Bot\Human differentiation of the English data. Nevertheless, the results are all in the same range at nearly 100% accuracy, which is a surprising outcome.

Moreover, it can be noticed, that the topic unigram feature is enough to enable a nearly perfect classification. The single feature accuracies (U) hardly differ from the results of the feature set (Fk) or the usage of all features (aF) in combination. Thus, the surface-level features and top 100 bigrams only minimally improve the accuracy in combination with the topic unigrams. With this knowledge the question arises, whether

 ISBN: 978-1-61208-891-4

the given data sets are possibly obtained or filtered for one or more specific topics. That would make the obtained results using topic unigrams less surprising. Unfortunately, there is no information available regarding the creation process.

The reason for the similarities between languages can be caused by their similar statistic characteristics or that the used approach is indeed language independent.

VI. Conclusion and Future Work

In this paper, topics as a feature for the author profiling classification tasks of differentiating between Twitter users and bots, as well as females and males was tested on a PAN data set containing English and Spanish Twitter data and found to surpass the results of existing works. Topics as feature were not considered in previous work. Furthermore, the tweets were first classified into the categories human and bot and the latter then further divided into female and male.

In summary, it was established that for the given author profiling tasks the topic feature in combination with a linear SVM provides the best results with accuracies up to 100%. This feature outperforms all other considered features except of bigrams, which yields similar performance.

Nevertheless, some improvements can be made in future works.

A second validation using a new completely independent data set would be useful. This data set should be created for English and for Spanish tweets without any topic restrictions. With this new data set, the overfitting hypothesis could be validated.

References

[1] I. Zeifman, "Bot Traffic Report 2016," Jan. 2017, [Last Accessed: 06-26-2021]. [Online]. Available: https://www.imperva.com/blog/bot-traffic-report-2016/?redirect=Incapsula

[2] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?" *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 6, pp. 811–824, 2012.

[3] S. Wojcik, S. Messing, A. Smith, L. Rainie, and P. Hitlin, "Bots in the Twittersphere," Apr. 2018, [Last Accessed: 05-16-2021]. [Online]. Available: https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/

[4] I. Vogel and P. Jiang, "Bot and Gender Identification in Twitter using Word and Character N-Grams," *Notebook for PAN at CLEF 2019*, 2019.

[5] A. Bessi and E. Ferrara, "Social bots distort the 2016 U.S. Presidential election online discussion," *First Monday*, vol. 21, no. 11, Nov. 2016. [Online]. Available: https://firstmonday.org/ojs/index.php/fm/article/view/7090

[6] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended Accounts in Retrospect: An Analysis of Twitter Spam," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 243–258. [Online]. Available: https://doi.org/10.1145/2068816.2068840

[7] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The Rise of Social Bots," *Commun. ACM*, vol. 59, no. 7, p. 96–104, Jun. 2016. [Online]. Available: https://doi.org/10.1145/2818717

[8] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches, "Overview of the author profiling task at pan 2013," *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pp. 352–365, 2013.

[9] PAN, "Offizielle PAN Aufgabenstellung: Bots and Gender Profiling 2019," [Last Accessed: 11-18-2020]. [Online]. Available: https://pan.webis.de/clef19/pan19-web/author-profiling.html

[10] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, "Online Human-Bot Interactions: Detection, Estimation, and Characterization," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 280–289, May 2017. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14871

[11] M. Martinc, I. Škrjanec, K. Zupan, and S. Pollak, "PAN 2017: Author Profiling-Gender and Language Variety Prediction (Notebook for PAN at CLEF 2017, 2nd place)," *Notebook for PAN at CLEF 2017*, 02 2018.

[12] M. Martinc, B. Škrlj, and S. Pollak, "Fake or Not: Distinguishing Between Bots, Males and Females (Notebook for PAN at CLEF 2019)," *Notebook for PAN at CLEF 2019*, 2019.

[13] S. Qi, L. AlKulaib, and D. A. Broniatowski, "Detecting and Characterizing Bot-Like Behavior on Twitter," in *Social, Cultural, and Behavioral Modeling*, R. Thomson, C. Dancy, A. Hyder, and H. Bisgin, Eds. Cham: Springer International Publishing, 2018, pp. 228–232.

[14] A. Basile *et al.*, "N-GrAM: New Groningen Author-profiling Model," *Notebook for PAN at CLEF 2017*, 2017.

[15] J. Pizarro, "Using N-grams to detect Bots on Twitter," in *CLEF 2019 Labs and Workshops, Notebook Papers*, L. Cappellato, N. Ferro, D. Losada, and H. Müller, Eds. CEUR-WS.org, Sep. 2019. [Online]. Available: http://ceur-ws.org/Vol-2380/

[16] R. F. S. Dias and I. Paraboni, "Combined CNN+RNN Bot and Gender Profiling," *Notebook for PAN at CLEF 2019*, 2019.

[17] J. P. Dickerson, V. Kagan, and V. S. Subrahmanian, "Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?" in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 2014, pp. 620–627.

[18] M.-C. Yang and H.-C. Rim, "Identifying interesting Twitter contents using topical analysis," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4330–4336, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414000141

[19] K. H. Lim, S. Karunasekera, and A. Harwood, "ClusTop: A Clustering-based Topic Modelling Algorithm for Twitter using Word Networks," in *2017 IEEE International Conference on Big Data (BIGDATA)*, 12 2017, pp. 2009–2018.

[20] A. Mukherjee and B. Liu, "Improving Gender Classification of Blog Authors," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, Oct. 2010, pp. 207–217. [Online]. Available: https://www.aclweb.org/anthology/D10-1021

 ISBN: 978-1-61208-891-4