Proceedings of the Workshop on Social Threats in Online Conversations: Understanding and Management (STOC-2020), pages 35–40
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Is this hotel review truthful or deceptive? A platform for disinformation detection through computational stylometry

Antonio Pascucci[1], Raffaele Manna[1], Ciro Caterino[2], Vincenzo Masucci[2], Johanna Monti[1]
L'Orientale University of Naples - UNIOR NLP Research Group[1], Expert System Corp.[2]
Via Duomo 219 Naples (Italy)[1], Via Nuova Poggioreale 60 Naples (Italy)[2]
{apascucci,rmanna,jmonti}@unior.it, {ccaterino,vmasucci}@expertsystem.com

Abstract

In this paper, we present a web service platform for disinformation detection in hotel reviews written in English. The platform relies on a hybrid approach of computational stylometry techniques, machine learning and linguistic rules written using COGITO©, Expert System Corp.'s semantic intelligence software thanks to which it is possible to analyze texts and extract all their characteristics. We carried out a research experiment on the *Deceptive Opinion Spam* corpus, a balanced corpus composed of 1,600 hotel reviews of 20 Chicago hotels split into four datasets: positive truthful, negative truthful, positive deceptive and negative deceptive reviews. We investigated four different classifiers and we detected that Simple Logistic is the most performing algorithm for this type of classification.

Keywords: Computational Stylometry, Disinformation Detection, Web Services.

1. Introduction

Disinformation is a phenomenon that is becoming part of everyday life. The phenomenon is uncontrollable, especially if we consider that social media and blogs are breeding grounds for news diffusion and that the higher the number of sharing of news, the more people are reached by the news. One of the fields in which disinformation is increasing quickly is hotel reviews, both for positive and for negative reviews. There may be an interest to spread positive or negative fake news about hotels. The main idea of our research is to reduce the impact of disinformation. For this reason, we developed a platform able answer to the question: is this hotel review truthful or deceptive? The paper is organized as follows: In Section 2 we present Related Work. In Section 3 we describe Computational Stylometry and some stylistic features and in Section 4 we present the *Deceptive Opinion Spam* corpus and we show the results of our testing. In Section 5 we propose the platform, ethical considerations are in Section 6 and Conclusions are in Section 7 along with Future Work.

2. Related Work

The proposed approach to detect disinformation in hotel reviews is certainly not the first one based on Computational Stylometry (CS) and Machine Learning (ML) / Deep Learning (DL) techniques. CS is presented in Section 4, DL exploits artificial neural networks with representation learning, while ML is the computer ability to learn from data. ML algorithms allow the system to preserve in its knowledge base each feature characteristic learned during the training process.

Despite "disinformation" and "fake news" represent two different concepts, they should be considered as close together, since they are both characterized by stylistic features typical of those who are lying. Disinformation is defined as *false information spread to deceive people*[1], while "fake news" *describes false stories that appear to be news, usually created to influence political views or as a joke*[2]. There is also a subtle difference between disinformation (incorrect information disseminated deliberately) and misinformation (that represents incorrect information disseminated unintentionally) (Egelhofer and Lecheler, 2019). (Kumar et al., 2016) investigated hoax articles presence on Wikipedia. The scholars used a large dataset of discovered hoaxes and detected that despite the community is efficient at identifying hoaxes, there is still a small number of these that survive for a long time. In their research, the scholars focused on the structure and content of the article and its mention in other articles. Their hoax/non-hoax classifier achieved an accuracy of 86% outperforming humans by a large margin (66%). In 2018, (Bakir and McStay, 2018) investigated the disinformation issue in the 2016 US presidential election campaign from an economic point of view. The scholars discovered a new version of disinformation, driven by profit and exploited by professional persuaders: it's about *emphatic media* (McStay, 2016), that represents personally and targeted news produced by *algo-journalism* (automated journalism), namely news articles generated by software through artificial intelligence.

As stated by (Lazer et al., 2018), addressing fake news requires a multidisciplinary effort. Despite authors of fake hotel reviews decide which words use, they can't handle the stylistic features that belong to the writing style and that make them unique. Considering that we detected stylistic features that characterize fake hotel review, we answer to (Lazer et al., 2018)'s request and we offer the potential of CS techniques in detecting fake hotel reviews. The "opinion spam" concept is very close to that of disinformation and mainly concerns in intentionally writing

[1] https://dictionary.cambridge.org/dictionary/english/disinformation
[2] https://dictionary.cambridge.org/dictionary/english/fake-news

fake reviews to products, restaurants or hotel (as in our case). The research of (Jindal and Liu, 2008) reveals that there are three different categories of opinion spam:

- untruthful opinions (undeserving positive reviews to some target objects to promote them or malicious negative reviews to some other objects to damage their reputation);
- reviews on brands only (those that do not comment on the specific product, but only the brand);
- non-reviews (those that are not reviews because contain advertisements)

(Ott et al., 2011) built a corpus composed of 400 truthful and 400 deceptive hotel reviews and proved that while n-grams based models are the best approach in identifying deceptive hotel reviews (89% of accuracy), a combination approach using psycholinguistically-motivated features (such as the number of words, lexical diversity, the score of narrativity) and n-grams features can perform slightly better (89.8% of accuracy). (Feng et al., 2012) exploit a Support Vector Machine (SVM) algorithm (Cortes and Vapnik, 1995) to build a classifier. The scholars used the corpus of (Ott et al., 2011) and two additional corpora and based their research on syntactical and lexical features and analyzed the text data with decision trees (DL approach) and achieved 91,2% of accuracy. The performances achieved by (Feng et al., 2012) improved those of (Ott et al., 2011), and demonstrated how a large use of personal pronouns (*I*) and possessive adjective (*my*) characterize deceptive hotel reviews.
(Popat et al., 2017) assessed the credibility of claims based on the occurrence of assertive and factive verbs, hedges, implicative words, report verbs and discourse markers. (Horne and Adali, 2017) focused on writing style and complexity to differentiate real news from fake news. The scholars used the number of occurrences of part-of-speech tags, swearing and slang words, stop words, punctuation, and negation as stylistic features. As stated by (Conroy et al., 2015), one of the best intuition in fake news and disinformation detection is that of (Feng et al., 2012): a deceptive writer with no experience with an event or object (e.g., never visited the hotel in question) may include contradictions or omission of facts present in profiles on similar topics.

3. Computational Stylometry

CS is a research area of Computational Linguistics that uses statistic techniques to analyze the literary style (Zheng et al., 2006). These techniques, through automatic linguistic analysis of texts, allow us to find countless personality traits. Wincenty Lutosławski (1863–1954), the one who coined the term *stylometry*, compared the style to handwriting: "If handwriting can be so exactly determined as to afford certainty as to its identity, so also with style, since style is more personal and characteristic than handwriting" (Lutosławski, 1897).
We have to consider that despite a deceptive review is written with greater intention to label it as positive or negative, stylistic features are not intentional but unintentional and result from sociological factors (such as age, gender and education level) and psychological factors (that include personality, mental health and being a native speaker or not) (Daelemans, 2013). It means that authors of deceptive review can certainly decide which words use in their review, but it is equally true that they can't handle the stylistic features that belong to their writing style. We believe that deceptive texts contain specific stylistic features that differentiate them from those truthful.

3.1. Stylistic Features

Almost all approaches in detecting disinformation and opinion spam focus on bag-of-words and part-of-speech models. As argued by (Ren and Ji, 2019) also linguistic (the functional aspect of a text), psychological (social, emotional and cognitive aspects), personal (any references to work, religion, etc.) and spoken (fillers and agreement words) features have to be taken into account. Several stylistic features characterize writing style and distinguish two or more different styles. Here we report a short list of stylistic features: sentence length (Argamon et al., 2003), word length distributions (Zheng et al., 2006), punctuation (Baayen et al., 1996), use of function words (Mosteller and Wallace, 1963), vocabulary richness (De Vel et al., 2001), use of a specific class of verbs or adjectives, use of first/third person.
Concerning CS, it is important to stress that stylometric analysis must focus only on unintentional choices by the writer of a text. Here we list some of the features that characterise deceptive texts in the corpus we investigated: high use of adverbs, high use of common nouns, high use of inappropriate lowercase on characters, high use of may/might and intensifiers, low use of punctuation, lower readability index, rare use of foreign terms, and high use of to + infinitive.

4. Corpus Analysis

We investigated the *Deceptive Opinion Spam* corpus in order to use it as pilot for the platform. The corpus consists of truthful and deceptive hotel reviews of 20 Chicago hotels and contains 400 truthful positive reviews from *TripAdvisor* and 400 deceptive positive reviews from *Mechanical Turk* described in (Ott et al., 2011) in addition to 400 truthful negative reviews from *Expedia*, *Hotels.com*, *Orbitz*, *Priceline*, *TripAdvisor* and *Yelp* and 400 deceptive negative reviews from *Mechanical Turk* described in (Ott et al., 2013). Each dataset consists of 20 reviews for each of the 20 most popular Chicago hotels.

4.1. Workflow

Our workflow for stylistic features extraction consists in the following steps:

- I) *Linguistic Definition of Stylometric Features*: since each author operates grammatical choices when writing a text, we organize all the grammatical characteristics of the texts under study in a taxonomy to detect the authorial fingerprint based on the grammatical choices done. This first step is carried

out thanks to COGITO©, that allows us to write LR and to perform word-sense disambiguation;

- II) *Semantic Engine Development*: we train the semantic engine to extract the features from the analyzed texts. The semantic engine is implemented thanks to COGITO©'s semantic network (*Sensigrafo*) - that can operate word-sense disambiguation - with the addition of the rules we built;

- III) *Training Set Analysis*: the training set is analysed and all features (based on the grammatical choices done by the writer) are extracted;

- IV) *ML*: In the last step, we exploit the features extracted to train the model to detect these features in the dataset. ML process is carried out exploiting WEKA platform (Hall et al., 2009) (a software with machine learning tools and algorithms for data analysis) and we build each classifier with the support of one of the algorithms available in WEKA.

4.2. Test

We built four classifiers trained with four different algorithms: Simple Logistic (SLO), Logistic (LOG), Sequential Minimal Optimization (SMO), and Random Forest (RFO). As we have mentioned, the whole corpus is composed of 1,600 reviews.
We decided to test all the aforementioned algorithms using the 10-folds cross-validation method. In Table 1 we show the 10-folds cross-validation results.

	SLO	LOG	SMO	RFO
10-f. cross-validation	**0,742**	0,721	0,738	0,702

Table 1: Percentage of correctly classified instances

Then, in order to evaluate the real performances of all the classifiers, we split the data into two sets: a training set composed of 1,200 of the 1,600 reviews and a test set composed of the remaining 400 reviews (200 truthful reviews and 200 deceptive randomly selected). In Table 2 we show the results of the test set.
According to Table 2 and to the confusion matrices in Figures 1, 2, 3, and 4, Simple Logistic is the best performing algorithm for this type of experiment and we decided to use it for our platform.

	SLO	LOG	SMO	RFO
Test experiment	**0,755**	0,707	0,725	0,710

Table 2: Percentage of correctly classified instances

The results we achieved (77.5%) do not improve those of (Ott et al., 2011) (89%) and those of (Feng et al., 2012) (89.9%). The reason is in the approach we adopted, that mainly focus on linguistic features and does not consider features (such as n-grams) that proved to be very useful in building deceptive detection models.

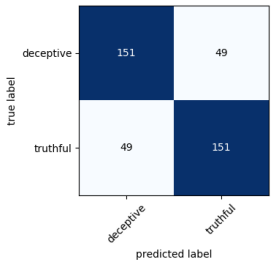

Figure 1: Confusion matrix of Simple Logistic classifier

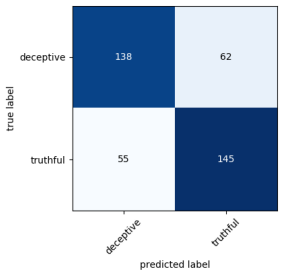

Figure 2: Confusion matrix of Logistic classifier

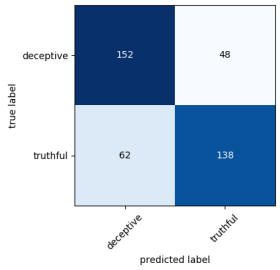

Figure 3: Confusion matrix of Sequential Minimal Optimization classifier

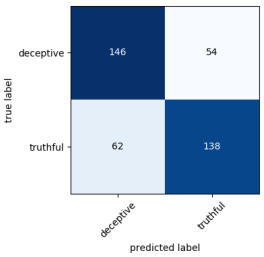

Figure 4: Confusion matrix of Random Forest classifier

5. Web service platform

The deceptive classification is provided through a REST web service which accepts as body input the text to classify.

The logic of the system consists of three main functional blocks:

- I) **Document Repository** - any document submitted to the system can be memorized together with a set of metadata about the document;
- II) **Computational Stylometry** - any document has to undergo a process of stylometric analysis. Thanks to our semantic intelligence software we can extract all stylistic features. The output is a set of stylometric features that are added to the document metadata (this block represents the whole workflow we have shown in Section 4.2, with the exception of ML process that is part of the third block);
- III) **Traits Prediction** - traits prediction refers to the profiling task thanks to ML techniques.

In Figure 5 we show the process. The method is POST, namely a method that accepts a text in the body and returns a JSON.

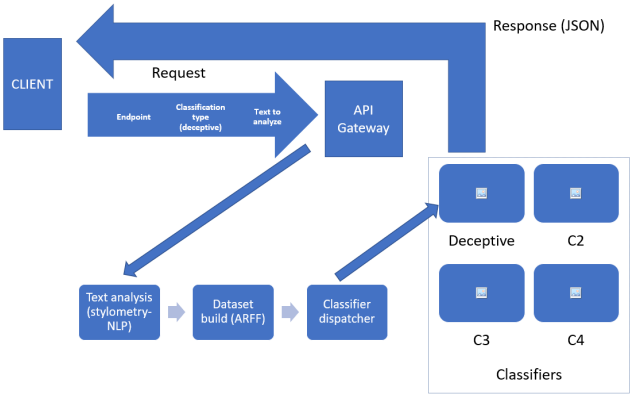

Figure 5: Web service platform process

The endpoint path contains information about the required type of classification, in this case, the *deceptive* one, so at the beginning, the user asks for a deceptive type of classification (it means the type of classification that the user needs and this type of classification includes two classes: *truthful* and *deceptive*) on the text the user provides. The API Gateway is in charge to receive requests and to begin the analysis process. The first step is the text analysis performed by the NLP technology, in order to extract stylometric features that will be used to classify the document in question. The second step is the ML process. For this process, we rely on WEKA platform (Hall et al., 2009), which requires a special file (ARFF) that contains all the information related to the text (namely the input text and the stylometric features extracted). The ARFF file is the input for the classifier, invocated from our classifier dispatcher module. Classification results are formatted in JSON and sent to the requester (CLIENT). Here we report an example of classification done on a text that belongs to the corpus:

Text:
After recent work stay at the Affinia Hotel, I can definitely say I will be coming back. They offer so many in room amenities and services, just a very comfortable and relaxed place to be. My most enjoyable experience at the Affinia Hotel was the amazing customization they offered, I would recommend Affinia hotel to anyone looking for a nice place to stay.

Prediction:
"actual": null,
"distribution":[0.7724841302455, 0.22751586975446],
"predicted": "deceptive",
"probability": 0.7724841302455,
"doc_name": "hotelopinion578-test"

The example reported above confirms that deceptive reviews are characterized by the use of intensifiers (*definitely*, *so many*, *a very*, *most enjoyable*). The review also lacks details, with reference only to general characteristics. Another characteristic that belongs to deceptive reviews is the repetition of the hotel name.
On these bases, our platform accepts hotel reviews written in English and returns to the user a prediction on the reliability of the review. It is important to stress, as shown in the example above, that the user receives a JSON that contains also a degree of probability of the prediction. Given the results of the test carried out on the *Deceptive Opinion Spam* we believe that our platform could make an important contribution to disinformation detection.

6. Ethical Considerations

The ethical argument has fundamental importance, especially if it is about public data closely linked to people. In fact, when we talk about author profiling and authorship attribution (two important branches of CS), we immediately think about the effects of our prediction. Then, privacy is the most important issue when we deal with profiling. In a case like this, we just need texts. All the other information (name of the authors, their age, thier origin and so on) are unnecessary. It means that possible negative impacts of our technology (the disinformation detection platform) are strongly mitigated. In other words, in the case of disinformation detection, it is not essential to know who wrote the review, and anonymization of reviews can mitigate ethical issues that may arise when these type of technologies are available to everyone.

7. Conclusions and Future Work

In this paper we have shown an experiment carried out on the *Deceptive Opinion Spam* Corpus, a corpus composed of 1,600 hotel reviews of 20 Chicago hotels split into four datasets: positive truthful, negative truthful, positive deceptive and negative deceptive reviews. The test has shown that the most performing algorithm is Simple Logistic, that correctly classified 75,5% of the test set we used. On the basis of these results, we developed a disinformation detection platform for hotel reviews written in English, in order to allow the user to submit a review and detect if it is deceptive or truthful and the percentage of probability of the prediction. It is not excluded that we will provide versions for other languages too. In this paper, we have shown how a linguistic-rule based approach can

help detect deceptive hotel reviews with good results. As a next step of our research we also aim to investigate more innovative techniques such as the use of neural networks and unsupervised learning approaches and to compare it with our current approach.

8. Acknowledgements

This research has been partly supported by the PON Ricerca e Innovazione 2014-20 and the POR Campania FSE 2014-2020 funds. Authorship contribution is as follows: Antonio Pascucci is author of Sections 1, 2, 3, 4, and 5. Sections 6 and 7 are in common between Antonio Pascucci and Raffaele Manna. This research has been developed in the framework of two Innovative Industrial PhD projects in Computational Stylometry (CS) by "L'Orientale" University of Naples in cooperation with Expert System Corp. We sincerely thank Ciro Caterino for helping us in developing the web service platform. We are also grateful to Vincenzo Masucci and Expert System Corp. for providing COGITO© for research and to Prof. Johanna Monti for supervising the research.

9. Bibliographical References

Argamon, S., Šarić, M., and Stein, S. S. (2003). Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480. ACM.

Baayen, H., Van Halteren, H., and Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Bakir, V. and McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital journalism*, 6(2):154–175.

Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Daelemans, W. (2013). Explanation in computational stylometry. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 451–462. Springer.

De Vel, O., Anderson, A., Corney, M., and Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64.

Egelhofer, J. L. and Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: a framework and research agenda. *Annals of the International Communication Association*, 43(2):97–116.

Feng, S., Banerjee, R., and Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Horne, B. D. and Adali, S. (2017). This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Eleventh International AAAI Conference on Web and Social Media*.

Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230.

Kumar, S., West, R., and Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lutosławski, W. (1897). *The origin and growth of Plato's logic: with an account of Plato's style and of the chronology of his writings*. Longmans, Green and Company.

McStay, A. (2016). Empathic media: The rise of emotion ai. *Arts & Humanities Research Council*.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 309–319. Association for Computational Linguistics.

Popat, K., Mukherjee, S., Strötgen, J., and Weikum, G. (2017). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012.

Ren, Y. and Ji, D. (2019). Learning to detect deceptive opinion spam: A survey. *IEEE Access*, 7:42934–42945.

Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American society for information science and technology*, 57(3):378–393.

10. Language Resource References

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 309–319. Association for Computational Linguistics.

Ott, M., Cardie, C., and Hancock, J. T. (2013). Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of*

the association for computational linguistics: human language technologies, pages 497–501.