Fact Checking or Psycholinguistics: How to Distinguish Fake and True Claims?

Aleksander Wawer, Grzegorz Wojdyga
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warszawa, Poland
{axw,g.wojdyga}@ipipan.waw.pl

Justyna Sarzyńska-Wawer
Institute of Psychology
Polish Academy of Sciences
ul. Jaracza 1
00-378 Warszawa, Poland
jsarzynska@psych.pan.pl

Abstract

The goal of our paper is to compare psycholinguistic text features with fact checking approaches to distinguish lies from true statements. We examine both methods using data from a large ongoing study on deception and deception detection covering a mixture of factual and opinionated topics that polarize public opinion. We conclude that fact checking approaches based on Wikipedia are too limited for this task, as only a few percent of sentences from our study has enough evidence to become supported or refuted. Psycholinguistic features turn out to outperform both fact checking and human baselines, but the accuracy is not high. Overall, it appears that deception detection applicable to less-than-obvious topics is a difficult task and a problem to be solved.

1 Introduction

Is deception detection more about writing style than verification of veracity against a database of credible information? Our paper attempts to answer this question by comparing approaches based on psycholinguistics with state-of-the-art fact checking systems.

In the case of the first method, the information is based on measuring psycholinguistic dimensions of language such as sentiment and emotional vocabulary, abstract or concrete character of utterances, analytical thinking, cognitive processes and so on. Using this type of features may lead to possibly more universal character of deception detection. According to Newman (Newman et al., 2003), the language of deception is linked to several psycholinguistic characteristics such as higher levels of abstraction. Psycholinguistic features were successful in the detection of falsified reviews (Ott et al., 2011) or prisoners lies (Bond and Lee, 2005). This method is universal and simple as no additional resources or references are necessary.

The second type of methods, namely fact checking systems, verify information using evidence from some credible source such as Wikipedia. Given a factual claim involving one or more entities (resolvable to Wikipedia pages), the system of this type must extract textual evidence (sets of sentences from Wikipedia pages) that support or refute the claim. Using this evidence, label the claim as supported, refuted (given the evidence) or not enough info if there isn't sufficient evidence. A number of systems of this type participated in Fever shared task (Thorne et al., 2018a).

2 Dataset

We analyzed 408 statements from 204 subjects who participated in a study of deception and deception detection conducted in the Institute of Psychology, Polish Academy of Sciences. Each subject was first asked to complete a short questionnaire. Based on its results we determined which two out of 12 debatable topics (eg. the right to abortion, attitudes towards immigrants, the best polish footballer, vegetarianism) the respondent has a clearly defined position on. Next they were asked to generate four statements. Two of them (which focus on one topic) were expressed in face-to-face communication and recorded while the other two were written on a web form (computer mediated communication). One statement on particular topic always represents the subject's real position while the other presents an opposing viewpoint. Subjects were also asked several standardized questions while giving statements so that each one contains the same elements: their stance, arguments for that position, and the subject's personal experience. The type of the statement (TRUE or LIE) as well as its form (writ-

Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 7–12
Hong Kong, November 3, 2019. ©2019 Association for Computational Linguistics

ten or oral) were counterbalanced. In this paper only written statements were analyzed. The statements were first translated into English using Google Translate. After that we checked the quality of translations and manually corrected a few of them.

3 Psycholinguistic Analysis

In order to obtain psycholingusitic descriptions of each utterance we applied the General Inquirer (Stone et al., 1966) – a tool for text content analysis which provides a wide range of categories. It helps to characterize text by defining words in terms of sentiment, intensity, varying social and cognitive contexts. Word categories were collected from four different sources: the Harvard IV-4 dictionary and the Lasswell value dictionary (Lasswell and Namenwirth, 1969), several categories were constructed based on work of Semin and Fiedler on social cognition and language (Semin and Fiedler, 1988), finally, marker categories were adapted from Kelly and Stone work on word sense disambiguation (Kelly and Stone, 1975). The full list of categories along with their descriptions can be found on the General Inquirer's home page[1].

4 Fact Checking

For fact checking we used two selected top performing systems from the Fever competition (Thorne et al., 2018b). The idea of Fever is to verify a claim based on the content of Wikipedia. In consists of three subtasks – firstly, given a claim, system should choose Wikipedia articles that might be useful in verifying. Next, the system has to pick up to 5 sentences that are crucial for verification. Finally, the system must decide whether the selected sentences support the claim, refute it or don't provide enough information. Labels are same as in SNLI (Bowman et al., 2015) and MNLI corpora (Williams et al., 2017).

4.1 Augmenting Article Database

We have verified that all the topics (such as abortion, immigrants, football players) were present in the English Wikipedia available in the Wikipedia resources for Fever (Thorne et al., 2018a) except of two that were specific for Polish common discourse – the famous Polish fitness trainer and the most famous Polish pseudo doctor. Therefore, we have translated their web pages from Polish Wikipedia [23] into English and added them to the resources that Fever systems are searching in. All the links that were present on their pages were redirected to their corresponding webpages in English Wikipedia.

4.2 Domlin

The Domlin system was introduced for Fever 2019 competition [4]. To our knowledge, the official article hasn't been published yet, but this model is similar to the previously introduced system for fact checking by the same authors (Stammbach et al., 2019). For retrieval task it uses the module, that was introduced by team athene (Hanselowski et al., 2018) for Fever 2018. It uses Wikipedia library [5] that wraps the Wikipedia API [6] which finds articles which title overlaps with the noun phrases within the claim. For sentence retrieval Domlin system is using the hierarchical retrieval approach, which finds the first sentence that is an evidence to support or refute the claim, and next, using all outgoing links it finds second sentence that might be part of evidence. For recognizing textual entailment Domlin system fine-tunes BERT language representation model (Devlin et al., 2019).

claim	Since prehistoric times man has hunted and ate meat, which allowed him to survive in those conditions.
label	SUPPORTS
evidence	Humans have hunted and killed animals for meat since prehistoric times.
	Meat is animal flesh that is eaten as food.

Table 1: Example of a correct fact verification by the Domlin system.

[1] http://www.wjh.harvard.edu/~inquirer/homecat.htm

[2] https://pl.wikipedia.org/wiki/Ewa_Chodakowska

[3] https://pl.wikipedia.org/wiki/Jerzy_Zi%C4%99ba

[4] https://github.com/dominiksinsaarland/domlin_fever

[5] https://github.com/goldsmith/Wikipedia

[6] https://www.mediawiki.org/wiki/API:Main_page

4.2.1 Analysis of results

More than 95% of results was labelled as ”Not enough info”. With ”Supports” and ”Refutes” results we have noticed that system was behaving correctly only sometimes. It found proper evidences and correctly labelled many claims, e.g. supported ”Vaccines are the best method to prevent serious infectious diseases.” or ”Meat has nutritional values, primarily protein.” and refutes to ”In addition, knowledge about vaccines is largely unverified”. The example of properly supported claim by the Domlin system is in table 1. Sometimes it made mistakes (like refutes ”Burning coal is dangerous to health and the environment.” where evidences did not indicate any of this). But very often it tried to prove claims that were impossible to verify such as: ”I will give an example.”, ”Why?” or ”I have this thesis in support.”. Example of such an example is in Table 2.

claim	I will give an example.
label	SUPPORTS
evidence	The name example is reserved by the Internet Engineering Task Force (IETF) in RFC 2606 [...] as a domain name that may not be installed as a top-level domain in the Domain of the Internet.
	Elliot John Gleave [...] better known by his stage name Example is an English rapper singer songwriter and record producer signed to Epic Records and Sony Music.

Table 2: Fact-checking of an unverifiable statement by the Domlin system.

4.3 UNC

The UNC system was the winner of FEVER 2018 task (Nie et al., 2019). In this system authors introduced Neural Semantic Matching Network (NSMN) which is modified version of ESIM (Chen et al., 2016). The NSMN is the architecture of neural network that is used in all three subtasks (document retrieval, sentence selection and claim verification). The three homogeneous neural networks conduct these tasks using some other features such as Pageview frequency and WordNet.

claim	Robert Lewandowski is a great Polish player
label	SUPPORTS
evidence	Robert Lewandowski [..] is Polish professional footballer who plays as a striker for [...] Bayern Munich.
	[...] he moved to top-flight Lech Poznan, and was the top scorer in the league as they won the 2009

Table 3: Example of a correct fact verification by the UNC system.

4.3.1 Analysis of results

More than 90% of results was labelled as ”Not enough info”. We have noticed behaviour similar to Domlin system – there were some correctly labelled statements (like ”Vaccinations protect against diseases by the stimulation of the man’s immune system”, another example in table 3), some mistakes and many tries of unverifiable claims (such as ”This is not good”, ”I will not agree to this”, ”Amen”). Interesting example is in Table 4 – one could argue whether the evidence supports the claim, but our insight is that this claim is not verifiable in the first place.

claim	Everyone should have a choice.
label	REFUTES
evidence	Most people regard having choices as a good thing , though a severely limited or artificially restricted choice can lead to discomfort with choosing and possibly an unsatisfactory outcome.

Table 4: Fact-checking of an unverifiable statement by the UNC system.

4.4 Verifiability

Our examination of fact checking systems revealed that systems try to find evidences to support or refute claims, that cannot be verified. Sentences like:”I will give an example.”, ”This is not good.”, ”These values that should be important to every citizen” are general opinions and cannot be verified. They are, however, processed because systems can find there noun phrases that are present in the Wikipedia (e.g. ”Example” as English rapper, ”This is not” – the fifth track from their Machine, ”Every” – title in the Baronetage of England). It is not a flaw – they had specific trainset, so it is natural that they ”overfit” and they don’t deal perfectly with new data.

It might, however, point an interesting direction in evolution of fact-checking systems and tasks. If a final goal is a real-life application, hence verifying statements or information that appear in a public discourse, it is crucial to face a problem that was just presented. Our idea is to include verifiability to the system. There are already important scientific works on verifiability e.g. (Newell et al., 2017) and factuality e.g. (Lioma et al., 2016). Based on these works it is worth to consider a binary falsifiability criterion – to determine whether it is possible to prove that given claim is wrong, hence whether it is possible to verify this claim in the first place. The term ”falsifiability” is inspired by Karl Popper’s scientific epistemology [7]. We believe that sentence can be consider falsifiable if and only if it describes facts about real objects. It is also worth to notice that task on distinction between opinions and facts was the topic of SemEval 2019, Task 8A [8]. Adding data with unverifiable statements and adding recognition of falsifiability as pre-processing might significantly help fact-checking systems to work in real-life applications.

5 Results: Psycholinguistics

For each utterance, we used the General Inquirer in order to compute frequency vectors corresponding to each of 182 categories in the General Inquirer dictionary. The vectors were then used as an input to supervised classification algorithms: Logistic Regression, Support Vector Machines with radial basis kernel (rbf), and XGBoost (Chen and Guestrin, 2016). We tested two variants of the feature space: with scaling (frequency as a percentage of a given category of words in all words) and raw word category frequencies. Table 5 contains mean accuracy of 20-fold cross-validation using each feature space variant. It reveals that the best performing classifier is XG Boost on scaled feature space, reaching 0.63 accuracy.

	scaled	raw
Logistic Regression	0.58	0.61
SVM (rbf)	0.57	0.60
XG Boost	0.63	0.59

Table 5: Mean accuracies of predicting deception in 20-fold cross-validation from the General Inquirer feature vectors.

6 Results: Fact Checking

In our experiments, we used each sentence of every utterance in our dataset as a claim to check with both Wikipedia-based fact checking engines (Fever shared task participants). We divided utterances to sentences using spaCy library [9]. Typically, most utterances contain between 5 and 15 sentences. Table 6 illustrates frequencies of labels generated by both tested systems represented as percentages.

	domlin	**unc**
NOT ENOUGH INFO	97.01%	93.84%
SUPPORTS	1.95%	4.21%
REFUTES	1.04%	1.95%

Table 6: Label percentages for both tested fact checking systems.

As it has been demonstrated, vast majority of sentences could not be fact-checked. However, for those that could, one may wonder how supported or refuted sentences predict honest (TRUE) or deceptive (LIE) utterances. We answer that question in Table 7 which shows the quality of such predictions on our data set as counts of each class as well as an overall accuracy.

7 Discussion

None of the tested methods achieved high accuracy. However, the problem is a very difficult one even for humans: it is well known and documented that most people perform poorly in lie detection experiments (Weinberger, 2010). Meta-analysis found that average accuracy in deception detection experiments is only 0.54, where 0.50 could

[7] ”I shall require that [the] logical form [of the theory] shall be such that it can be singled out, by means of empirical tests, in a negative sense: it must be possible for an empirical scientific system to be refuted by experience.” *The Logic of Scientific Discovery*

[8] https://competitions.codalab.org/competitions/20022

[9] https://spacy.io/

	domlin	**unc**
SUPPORTS-LIE	22	63
REFUTES-LIE	14	26
SUPPORTS-TRUE	26	58
REFUTES-TRUE	10	30
ACCURACY	0.55	0.47

Table 7: Label percentages for both tested fact checking systems.

be obtained by chance. This finding is extremely stable, with 90% of published studies reporting results within 0.1 of the across-study mean(Bond Jr and DePaulo, 2006). Studies show also that there is very little variance attributable to individual differences in judge ability (Bond Jr and DePaulo, 2008) or judge professional experience ((Aamodt and Custer, 2006), (Bond Jr and DePaulo, 2006)).

In the context of such baselines, one should not consider the results obtained using pschycholinguistic text features as entirely discouraging. The best of tested methods (XG Boost classifier) achieved mean accuracy of 0.63.

Using Wikipedia information to verify the veracity of utterances is not particularly useful when applied to a dataset of opinionated, often polarizing topics such as vegetarianism and abortion. This may be due to several factors. First, Wikipedia, as a community-edited resource, may simply not contain controversial or debatable claims. Second, lying seems to be a broad phenomenon, referring to the experiences, feelings and opinions of a given person and related to both cognitive and emotional load, which may end up in not referring to verifiable facts.

8 Conclusion and Future Work

In this paper we have compared two approaches to deception detection: fact checking and psycholinguistic features. We used data from a large ongoing study on deception detection in Polish. We concluded that psycholinguistic approach has an advantage, but the results may be related to often opinionated and controversial topics covered in the study, not easy for fact checking systems based on Wikipedia. The problem not only in very low recall (majority of sentences labelled as not enough info) but also in low precision when predicting deceptive utterances. In order to make our findings more broad, we plan to apply the same approach to other data types such as fake news.

9 Acknowledgements

Research funded by the National Science Centre Poland grant UMO-2017/26/D/HS6/00212

References

Michael G Aamodt and Heather Custer. 2006. Who can best catch a liar? *Forensic Examiner*, 15(1).

Gary D Bond and Adrienne Y Lee. 2005. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, 19(3):313–329.

Charles F Bond Jr and Bella M DePaulo. 2006. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234.

Charles F Bond Jr and Bella M DePaulo. 2008. Individual differences in judging deception: Accuracy and bias. *Psychological bulletin*, 134(4):477.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv preprint arXiv:1809.01479*.

Edward F Kelly and Philip J Stone. 1975. *Computer recognition of English word senses*, volume 13. North-Holland.

Harold D Lasswell and J Zvi Namenwirth. 1969. The lasswell value dictionary. *New Haven*.

Christina Lioma, Birger Larsen, Wei Lu, and Yong Huang. 2016. A study of factuality, objectivity and relevance: three desiderata in large-scale information retrieval? In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, pages 107–117. ACM.

Edward Newell, Ariane Schang, Drew Margolin, and Derek Ruths. 2017. Assessing the verifiability of attributions in news text. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 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.

Gün R Semin and Klaus Fiedler. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of personality and Social Psychology*, 54(4):558.

Dominik Stammbach, Stalin Varanasi, and Günter Neumann. 2019. Domlin at semeval-2019 task 8: Automated fact checking exploiting ratings in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1149–1154.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The Fact Extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*.

Sharon Weinberger. 2010. Airport security: Intent to deceive? *Nature News*, 465(7297):412–415.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.