Information Nutritional Label and Word Embedding to Estimate Information Check-Worthiness

Cédric Lespagnol
IRIT, URM5505 CNRS
UPS,Université de Toulouse
Toulouse, France
cedric.lespagnol@univ-tlse3.fr

Josiane Mothe
IRIT, URM5505 CNRS
ESPE, Université de Toulouse
Toulouse, France
josiane.mothe@irit.fr

Md Zia Ullah
IRIT, UMR5505 CNRS
UPS, Université de Toulouse
Toulouse, France
mdzia.ullah@irit.fr

ABSTRACT

Automatic fact-checking is an important challenge nowadays since anyone can write about anything and spread it in social media, no matter the information quality. In this paper, we revisit the information check-worthiness problem and propose a method that combines the "information nutritional label" features with POS-tags and word-embedding representations. To predict the information check-worthy claim, we train a machine learning model based on these features. We experiment and evaluate the proposed approach on the CheckThat! CLEF 2018 collection. The experimental result shows that our model that combines information nutritional label and word-embedding features outperforms the baselines and the official participants' runs of CheckThat! 2018 challenge.

CCS CONCEPTS

• **Information systems → Information retrieval**; *Content analysis and feature selection*; *Clustering and classification*;

KEYWORDS

Information retrieval, Information check-worthiness, Fact-checking, Information nutritional label, CLEF evaluation forum

ACM Reference Format:
Cédric Lespagnol, Josiane Mothe, and Md Zia Ullah. 2019. Information Nutritional Label and Word Embedding to Estimate Information Check-Worthiness. In *42nd Int'l ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331298

1 INTRODUCTION

Social media eases information spreading, makes information diffusion quicker, and reaches potentially more people than traditional media [11], in many cases regardless of the information quality. For example, Allcott and Gentzkow reported that the 115 pro-Trump fake stories the authors collected were shared 30 million times on Facebook [2].

Public authorities are now aware of the possible negative impact of fake news spreading. For example, in January 2018, the French president Emmanuel Macron "has vowed to introduce a law to ban fake news on the Internet during French election campaigns."[1] Although the concrete implementation of fact checking is very challenging, researchers have a major role to play to inform users and citizens in detecting which information needs to be checked.

Automate fact-checking has recently become a hot topic since it could be used to warn social media users and readers or even to stop the spreading of fake news. Moreover, automatic fact-checking has been the subject of challenges and tracks in evaluation forums such as SemEval 2017 shared task on Rumor Detection that "aims to identify and handle rumors and reactions to them, in text" [4]. Another ongoing challenge is FEVER (Fact Extraction and VERification)[2] at EMNLP 2019 which goal is "to evaluate the ability of a system to verify information using evidence from Wikipedia." CLEF 2018 also introduced the CheckThat! Lab on Automatic Identification and Verification of Claims in Political Debates [16] that "aims to foster the development of technology capable of both spotting and verifying check-worthy claims in political debates." This Lab consists of two challenges[3]: the first one is to predict which claims in political debates should be prioritized for fact-checking (i.e. check-worthy) and the second one is to automatically estimate the level of fact-checking of the check-worthy claims.

In this paper, we focus on the first challenge. The originality of our approach resides in the way we represent the information. It relies on the *Information Nutritional Label* for online documents [6]. This label was initially introduced to "help readers making more informed judgments about the items they read." The nutritional label provides scores for various criteria to describe the content of a written text. Although the information nutritional label aims at guiding the information reader only, we hypothesize that part of it could be used to decide whether a piece of information should be prioritized for checking or not.

We combine the features from the information nutritional label with POS-tags or word-embedding representations and learn a machine learning model to predict the information check-worthiness. Our results show that combining the nutritional label with word-embedding improves the performance compared to more standard, although quite recent approaches [9, 21]. In the evaluation part, we also analyze the impact of the different feature combinations and the performance of different machine learning algorithms.

The remaining of this paper is organized as follows: Section 2 presents the related work. Section 3 presents the different types of features used in our machine learning-based models. Experimental

SIGIR '19, July 21–25, 2019, Paris, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6172-9/19/07...$15.00
https://doi.org/10.1145/3331184.3331298

[1] https://www.theguardian.com/world/2018/jan/03/emmanuel-macron-ban-fake-news-french-president

[2] http://fever.ai/

[3] http://alt.qcri.org/clef2018-factcheck/

results and discussion are described in Section 4. Finally, Section 5 concludes this paper and shows some future perspectives.

2 RELATED WORK

Related work on automated fact-checking mainly focuses on verifying the veracity of claims; a few studies address the challenge of identifying check-worthy statements. One of such first studies was ClaimBuster [10] that extracts check-able claims, classifies, and ranks the check-worthiness of these claims. The authors used the transcripts of all of the 30 US presidential debates until 2012 that were manually annotated as Non-Factual Sentence (NFS), Unimportant Factual Sentence (UFS), or Check-worthy Factual Sentence (CFS). The authors used an SVM with sentence-level features such as sentiment, length, TF-IDF, POS-tags, and Entity Types. They achieved an average precision (AP) of 0.223 for the top-100 sentences. One possible improvement of the ClaimBuster system was to consider the context of the claim to evaluate. Gencheva et al. integrated several context-aware and sentence-level features to train both SVM and Feed-forward Neural Networks [7]. This approach outperforms the ClaimBuster system in terms of MAP and precision.CheckThat! Lab at CLEF 2018 is the most recent challenge on this topic. Several teams participated, including Prise de Fer [21], Copenhagen [9], bigIR [20], UPV-INAOE-Autoritas [8], and RNCC [1]. The best performing system is Prise de Fer [21] that obtained a MAP score of 0.133, outperforming the second best system [9] by 18%. The authors represented the sentence using word-embedding combined with POS-tags, syntactic dependencies, and some new features including named entities, sentiment, and verbal forms. With gathering external training data, this sentence representation was used as input to train a multi-layer perceptron (MLP) which is composed of two hidden layers (with 100 units and 8 units, respectively) and the hyperbolic tangent (tanh) as an activation function. The other participants used different representations such as character n-grams [8] or topics [20]. The participants used different machine learning (ML) algorithms such as SVM [1], Random Forest [1], k-nearest neighbors [8], or Gradient boosting [20]. In this paper, we investigate combining different types of features in the text representation, including some from the information nutritional label [6].

3 FROM NUTRITIONAL LABEL TO FEATURE VECTORS TO LEARN CHECK-WORTHINESS

3.1 Information nutritional label features

The Information Nutritional Label for Online Documents [6] aims at helping on-line information users in their information "consumption." The proposed label describes a textual information unit according to the following criteria: (1) **Factuality**: the number of facts it mentions, (2) *Readability*: the ease with which a reader can understand it, (3) *Virality*: the speed at which it is propagated, (4) **Emotion**: its emotional impact, (5) *Opinion*: the number of opinionated sentences it contains, (6) **Controversy**: the number of controversial issues it addresses, (7) *Authority/Trust/Credibility*: its credibility and the authority and trust of the source it belongs to, (8) **Technicality**: the number of technical issues it addresses and technical terms used, and (9) *Topicality*: its current interest which is time-dependent.

Although the information nutritional label aims at guiding the information reader only, we hypothesize that part of it could be used to decide whether a piece of information should be prioritized for checking or not. From the initial label, we have not used all the criteria since some are not straightforward applicable to the political transcribes that compose the data set we use (e.g. readability, topicality, or authority). We kept the ones that are described below.

3.1.1 *Factuality.*

We develop two feature variants for the factuality as follows:

- **Factuality_Proba** that computes the probability of a sentence to be representative of a fact.
- **Factuality_Strict** is a binary feature which considers a sentence as either a fact (1) or an opinion (0). This feature is 1 if Factuality_Proba is ≥ 0.5 and is 0 otherwise.

These two features are extracted using Matatusko's classifier[4]. It is based on a multi-layer perceptron (MLP) using LBFGS gradient descent [18]. To train the classifier, we collected the data from Wikipedia articles for factual sentences (World_War_I, Industrial_Revolution, October_Revolution, Fermi_paradox, Steam_engine, Barack_Obama, Amazon_(company), Netherlands, Triangular_trade, Song_dynasty, Nanking_Massacr, The_Holocaust) and from the Opinosis site `http://kavita-ganesan.com/opinosis` for opinion sentences. To represent a sentence as features, given each word of the sentence, we estimate the occurrence of POS tags, entity types, and dependency tags, adapted from the *spacy* annotation.

3.1.2 *Emotion.*

We hypothesize that a sentence with a high emotional level might be used to deceive the "consumer" towards accepting false information and thus that it should be checked. We use the list of 2,477 emotional words with evaluation from AFINN[5] [17], for example, abusive (-3), proud (2), etc. We develop three feature variants for the emotion criterion as follows:

- **Emotion_P** (resp. **Emotion_N**) is the sum of the positive (resp. negative) rating of the words in the sentence.
- **Emotion_U** considers both positive and negative emotions to get an overall level by summing the absolute value of the positive and negative rating of the words in the sentence.

3.1.3 *Controversy.*

We hypothesize that a sentence addressing a controversial issue is more likely to be worth checking. To estimate the controversy level of a sentence, we count the number of controversial issues in the sentence based on the controversial entries in Wikipedia article "Wikipedia:List_of_controversial_issues." For each controversial issue, we also take into account the anchor text labels to find the synonyms and other designations of the issues. Thus, we build a dictionary of the controversial noun phrases which is used to estimate the controversy level of the sentence.

3.1.4 *Technicality.*

We hypothesize that technical words will be associated less with check-worthiness. To estimate this criterion, we develop two feature variants by counting the number of domain-specific words found in a sentence in two different ways as follows:

- **Technicality_RE** uses a specific regular expression defined in [12] to find domain-specific words. First, we use NLTK [3] for POS tagging; then, we identify the terminological noun phrases

[4] `https://github.com/matatusko/opinion-or-fact-sentence-classifier`
[5] `http://www2.imm.dtu.dk/pubdb/p.php?6010`

(NPs) using the Python regular expression library. NPs represents domain-specific words. We calculate the number of these NPs that occur more than once.

- **Technical_List** uses technical words from the Academic vocabulary Lists (`https://www.academicvocabulary.info`). There are about 8,000 words that occur at least three times more frequently than expected in one of the nine COCA-Academic domains (e.g. Law, Medicine or Technology). We count the number of technical words in the sentence. All the features are normalized by dividing the feature value by the sentence length.

3.2 Features on word embedding and spaCy

3.2.1 Word embedding. Word embedding refers to the representation of a word in semantic space as a vector of numerical values. Words that are semantically and syntactically similar tend to be close in this embedding space. "Word vectors" was trained on GoogleNews corpus using Word2Vec model [14]. We average the *word vectors* of every word in a sentence. When we could not find a word in the model, we represent it with a zero vector. Although zero vector affects the mean [19], this is indeed essential when we could not find any word of the sentence in the model.

3.2.2 SpaCy annotations. We use fine-grained POS tags, syntactic dependency tags, and the entities from a sentence as features using the information extraction library, spaCy[6]. To extract these features, we collect all the POS tags, dependency tags, and entity types mentioned in SpaCy. Then, for each word within a sentence, we measure the number of occurrences of all the collected tags. Simplier methods could also be used to extract key-phrases [15], but we keep this for future work.

4 EVALUATION

4.1 Data collection

We evaluated our model on the CLEF18 CheckThat! 2018 collection (CT-CWC-18) [16] which is composed of the transcriptions of political debates or speeches from the 2016 US Presidential campaign. The "CT-CWC-18" collection is divided into a training and test set. Each line in a transcription file consists of the line number (LN), the Speaker name, the transcription of the statement that the Speaker said, and for the training set a label indicating whether this statement is check-worthy (1) or not (0). The training set contains the Presidential debates (first and second) and the Vice-Presidential debate while the test set consists of the third Presidential and ninth Democratic debates along with five of Donald Trump's speeches. Table 1 shows some statistics of the training and testing sets [16].

4.2 Evaluation measures

We consider several evaluation measures including mean average precision (MAP) which was the official measure for the CLEF track [16] mean reciprocal rank (MRR), mean R-precision (MRP), and mean precision at 5 (MP@5). We use the scripts provided by the CheckThat! Lab organizers [7].

Table 1: CT-CWC-18 collection: number of sentences (#Sent.) and the number of check-worthiness (#CW) on the training and test sets.

	Type	**Set**	**#Sent.**	**#CW**
Debates	First Presidential	Train	1,403	37
	Second Presidential	Train	1,303	25
	Vice-Presidential	Train	1,358	28
	Third Presidential	Test	1,351	77
	Ninth Democratic	Test	1,464	17
Speeches	Donald Trump Acceptance	Test	375	21
	Trump at the World Economic Forum	Test	245	11
	Trump at a Tax Reform Event	Test	412	16
	Trump's Address to Congress	Test	390	15
	Trump's Miami Speech	Test	645	35
	Total		**8,946**	**282**

4.3 Experiments and Results

4.3.1 Considering various combination of features. To represent a sentence, we considered three categories of features, namely information nutritional label (denoted as N), spaCy annotations (S), and word-embedding (W). Since some features may be complementary, we explore the combinations of features which could address the check-worthiness claim. We fed these categories of features one by one and also every possible combination to the ML algorithms for training the models. The considered ML algorithms includes Random Forest (RF), SVM_RBF, SVM_Linear, MLP_LBFGS (one hidden layer with 100 units, Relu activation function, and optimizing "log" loss function using LBFGS), and SGD_Logloss (Stochastic gradient descent classifier training using "log" loss function, AKA, Logistic regression). Moreover, the distribution of the two classes (Check-worthy or not) in the training set is unbalanced (97%/3%) (see Table 1). To balance the distribution, we applied oversampling on the training set with the SMOTE algorithm [13].

Table 2: MAP of the ML algorithms considering different groups of features (Nutritional label (N), SpaCy (S), and Word-embedding (W)). The best MAP achieved by the top-ranked run from "Prise de Fer" team [21] at Checkthat! 2018 is 0.133. Our best result (0.230) is obtained without oversampling with SGD ML and "NW" features.

	MAP	**N**	**S**	**W**	**NS**	**NW**	**SW**	**NSW**
Oversampled	RF	.072	.097	.183	.103	.128	.130	.119
	SVM_RBF	.080	.117	.201	.110	.202	.131	.131
	SVM_Linear	.071	.102	.130	.112	.129	.135	.128
	MLP_LBFGS	.098	.101	.142	.118	.141	.119	.119
	SGD_Logloss	.065	.081	.146	.122	.152	.126	.129
Without	SGD_Logloss	.079	.049	.210	.108	**.230**	.099	.107
	MLP_LBFGS	.097	.086	.131	.086	**.183**	.116	.101
	SVM_Linear	.085	.118	.176	.114	**.180**	.129	.131
	SVM_RBF	.072	.092	**.172**	.094	.159	.111	.115
	RF	.066	.097	.089	.099	.084	**.091**	.087

Table 2 presents the comparative performance in terms of MAP considering the different combinations of features with and without

[6] `https://spacy.io/`

[7] `http://alt.qcri.org/clef2018-factcheck`

oversampling the training set. While it is not always the case, in most cases oversampling does not improve the results. Moreover, we can see that the best combination of features is based on information nutritional label with word-embedding ("NW") with which SGD_Logloss achieves the best MAP score of 0.230 without oversampling. It seems that some methods benefit from oversampling such as Random Forest (RF) and SVM_RBF; we keep a finer analysis of this phenomenon for future research. In the remaining of the experiment report, we focus on the non-oversampled training set.

4.3.2 Considering various ML algorithms. Given the best combination of features ("NW") explored in Section 4.3.1, we estimated different measures for various ML algorithms. We also compared with two baselines: the first baseline, N-GRAM is an SVM classifier with "RBF" kernel (C=10, γ=0.1) where each sentence is represented using uni-gram features. The second baseline is Random which scores each sentence from the test file using a random weight. We also compared with the best participants' methods in CLEF CheckThat! 2018 task1.

Table 3: MAP, MMR, MR-P, and MP@5 of "NW" combination of features compared to two baselines and best participants' methods at CLEF CheckThat! 2018. SGD_Logloss classifier outperforms any of the participants' runs on all the measures apart for MP@5 where the results are equal to the best participant's results.

	Method	**MAP**	**MRR**	**MR-P**	**MP@5**
BL	N-GRAM	.120	.409	.128	.171
	Random	.049	.063	.036	.000
Particip.	Prise de Fer	.133	.497	.135	.200
	Copenhagen	.115	.316	.110	.114
	UPV-INAOE	.113	.462	.132	**.314**
	bigIR	.112	.262	.117	.114
Without	SGD_Logloss	**.230**	.573	**.254**	**.314**
	MLP_LBFGS	.183	.391	.197	.257
	SVM_Linear	.180	**.626**	.164	.286
	SVM_RBF	.159	.422	.151	.286
	RF	.084	.125	.078	.086

From Table 3, we can see that SGD_Logloss using "NW" features consistently outperforms the participants' methods, the baselines, as well as our other ML variants using the same "NW" features and this is for the official measure MAP and on most of the other measures. SVM_Linear performs very well on the ranked based measure MRR while it does not on the other measures. We will investigate the reason for this in future deeper analysis.

5 CONCLUSIONS

We have proposed a method for predicting information check-worthiness using features based both on the information nutritional label and POS-tags or word-embedding representations. Experimental results on the CheckThat! 2018 collection shows that combing information nutritional label and word-embedding outperforms the baselines and the known related methods. Oversampling the training set have not improved the results although the training examples are unbalanced. We will focus on this issue as well as feature selection in future work. We also would like to study additional components from the information nutritional label such as readability for which many measures exist [5] and could be reused.

REFERENCES

[1] Romain Agez, Clément Bosc, Cédric Lespagnol, Noémie Petitcol, and Josiane Mothe. 2018. IRIT at CheckThat! 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France.*

[2] Hunt Allcott and Matthew Gentzkow. 2017. *Social media and fake news in the 2016 election.* Technical Report. National Bureau of Economic Research.

[3] Edward Loper Bird, Steven and Ewan Klein. 2009. Natural Language Processing with Python. (2009).

[4] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proc. of the 11th International Workshop on Semantic Evaluation.* ACL, 69–76.

[5] Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. 2019. A survey on evaluation of summarization methods. *Information Processing and Management* (2019). https://doi.org/10.1016/j.ipm.2019.04.001

[6] Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Jarvelin, Rosie Jones, YiquN Liu, Josiane Mothe, Wolfgang Nejdl, et al. 2018. An Information Nutritional Label for Online Documents. In *ACM SIGIR Forum*, Vol. 51. ACM, 46–66.

[7] Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017.* 267–276.

[8] Bilal Ghanem, Manuel Montes-y-Gómez, Francisco M. Rangel Pardo, and Paolo Rosso. 2018. UPV-INAOE - Check That: Preliminary Approach for Checking Worthiness of Claims. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France.*

[9] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, and Christina Lioma. 2018. The Copenhagen Team Participation in the Check-Worthiness Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 CheckThat! Lab. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France.*

[10] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. *world* (2015).

[11] Thi Bich Ngoc Hoang and Josiane Mothe. 2018. Predicting information diffusion on Twitter–Analysis of predictive features. *Journal of computational science* 28 (2018), 257–264.

[12] John S Justeson and Slava M Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering* 1, 1 (1995), 9–27.

[13] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26.* Curran Associates, Inc., 3111–3119.

[15] Josiane Mothe, Faneva Ramiandrisoa, and Michael Rasolomanana. 2018. Automatic keyphrase extraction using graph-based methods. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing.* ACM, 728–730.

[16] Preslav Nakov, Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. In *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF'18).* Springer, 372–387.

[17] F. Å. Nielsen. 2011. AFINN. www2.imm.dtu.dk/pubdb/p.php?6010

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[19] Md Zia Ullah, Md Shajalal, Abu Nowshed Chy, and Masaki Aono. 2016. Query subtopic mining exploiting word embedding for search result diversification. In *Asia Information Retrieval Symposium.* Springer, 308–314.

[20] Khaled Yasser, Mucahid Kutlu, and Tamer Elsayed. 2018. bigIR at CLEF 2018: Detection and Verification of Check-Worthy Political Claims. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France.*

[21] Chaoyuan Zuo, Ayla Karakas, and Ritwik Banerjee. 2018. A Hybrid Recognition System for Check-worthy Claims Using Heuristics and Supervised Learning. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France.*