LITL at SMM4H: an old-school feature-based classifier for identifying adverse effects in Tweets

Ludovic Tanguy, Lydia-Mai Ho-Dac, Cécile Fabre
and the **Master LITL students**: Roxane Bois, Touati Mohamed Yacine Haddad,
Claire Ibarboure, Marie Joyau, François Le moal, Jade Moillic,
Laura Roudaut, Mathilde Simounet, Irena Stankovic, Mickaela Vandewaetere

CLLE: CNRS & University of Toulouse, France
{ludovic.tanguy, lydia-mai.ho-dac, cecile.fabre}@univ-tlse2.fr

Abstract

This paper describes our participation to the SMM4H shared task 2. We designed a linear classifier that estimates whether a tweet mentions an adverse effect associated to a medication. Our system addresses English and French, and is based on a number of ad-hoc word lists and features. These cues were mostly obtained through an extensive corpus analysis of the provided training data. Different weighting schemes were tested (manually tuned or based on a logistic regression), the best one achieving a F1 score of 0.31 for English and 0.15 for French.

1 Overview

This article describes the participation of the students of the *LITL* master and their teachers to the Social Media Mining for Health (SMM4H) shared task 2 (Klein et al., 2020). LITL (stands for *Linguistique, Informatique, Technologies du Langage*, i.e. Linguistics, IT, Language technologies) is a master's program at the University of Toulouse, France that is mainly aimed at linguistics and humanities students.

The shared task is a binary classification of Twitter messages in different languages, indicating whether the message contains a mention of medication adverse effects. Participation to this task was part of the first year students' curriculum. At this stage, their computer skills were still limited to corpus processing and simple programs, so it was decided that the system's architecture would be a traditional linear classifier based on ad-hoc features. This approach was also deemed justified given the heavily biased distribution of data (known to be an issue for most machine learning techniques). However, the students were encouraged to apply and hone their corpus linguistics skills, and to perform some feature engineering. The approach was the following:

1. Observe the training data with corpus analysis tools, in order to identify the main characteristics of the target (i.e. tweets evoking an adverse effect);
2. Build word lists and design simple features for each of these characteristics;
3. Design a program that computes the features' values on the target data and implements a simple weight-based linear classifier;
4. Tune the weights in order to maximize the classifier's performance on the validation data.

Due to the necessity to actually observe and understand the training data only the French and English sets were considered, as none of the students was proficient in Russian.

2 Technical details

For observation and actual processing in both languages the tweets were preprocessed as follows:

- retweet marks (`rt @X`) were removed;
- user names (`@XXX`) were replaced with a generic and POS-wide unambiguous proper name (*Sacha*);
- URLs and email addresses were replaced with generic placeholders (`<URL/>` and `<email/>`);
- non-standard spelling was normalised (e.g. removal of exceeding repeated letters *baaad* → *bad*);
- POS tagging and lemmatizing were performed, using the Talismane toolkit for both target languages (Urieli, 2013).

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 134–137
Barcelona, Spain (Online), December 12, 2020.

Fifteen word lists were compiled for each language, each one targeting a specific aspect of the tweets content. Those word lists contain keywords extracted from the target tweets and non target tweets using the TXM corpus analysis tool (Heiden et al., 2010). The lists were extended with existing lexical resources such as sentiment lexicons (e.g. the SocialSent lexicon (Abdaoui et al., 2017) and the FEEL – French Expanded Emotion Lexicon (Hamilton et al., 2016)) and biomedical domain language resources (Névéol et al., 2014). Table 1 gives an overview of the word lists designed and used as features for English and French Tweet classification.

Target tweet keywords (i.e. positively correlated with adverse effect)			
Word list	**Example (English)**	**# items (English)**	**# items (French)**
Symptoms	*headache*, *cough*, *addict*...	378	376
Causal verbs	*impact*, *stop*...	41	58
Sentiment (negative)	*dirty*, *resent*...	3647	894
Body parts	*chest*, *joint*...	86	84
Medication (first set)	*Effexor*, *Paxil*...	21	22
Increase verbs	*gain*, *raise*...	52	112
Decrease verbs	*decline*, *reduce*...	38	107
First person pronouns and determiners	*I*, *our*	7	12
Negation	*not*, *cannot*...	9	15
Emojis (negative)	🙁, 😵	23	23
Non target tweet keywords (i.e. negatively correlated with adverse effect)			
Word list	**Example (English)**	**# items (English)**	**# items (French)**
Misc. verbs	*approve*, *study*...	18	8
Sentiments (positive)	*great*, *secure*...	2080	848
Medication (second set)	*Floxin*, *Prozac*...	40	9
2nd and 3rd person pronouns and determiners	*you*, *himself*	17	22
Emojis (positive)	🙂	57	57

Table 1: Word lists designed and used as features

Each word list led to a numeric feature corresponding to the raw frequency of matching lemmas in the tweet. Two different strategies were considered for dealing with multi-word expressions. Runs 1 and 3 count all items as matches even in case they are also part of an item in another list, e.g. *skin* (body part) and *skin rash* (symptom). In contrast, run 2 only counts the longer item (e.g. *skin rash* as a symptom feature).

Three additional non-lexical features were also used: number of hash tags, number of URLs and number of Twitter user names (i.e. *Sacha*, cf. *supra*).

Each feature was assigned a weight proportional to its relative importance in the decision process. For runs 2 and 3 the weights were individually fixed based on the frequency ratios in target (*vs* non target) tweets in the training data, and then manually adjusted based on the scores obtained on the validation sets. The best weights were found by progressively increasing the weight of each feature independently of each other until the best F1 score is reached. For run 2, we adopted a principle of equality between features. For run 3, some features were considered as more important than others on the basis of manual observations. Run 1 used a standard logistic regression classifier trained on training data.

3 Results and discussion

Table 2 shows the results for each run and for each language on the validation and test sets. The first strategy for dealing with multi-word expressions was clearly better. Manual tuning of the feature weights (which was performed before the students were introduced to machine learning techniques) was promising on the validation set (especially regarding precision) but proved to be much less robust in the test set. Further experiments will be performed in order to assess the added value of selected word lists, compared to more straightforward and non-selective bag-of-words methods, and of course more recent NLP techniques based on word embeddings and neural classifiers.

Language	Run	Validation			Test
		R	P	F1	F1
English	1 (logistic regression, all items)	0.40	0.16	0.23	**0.31**
	2 (adjusted, longer items only)	**0.55**	0.23	0.32	0.25
	3 (adjusted, all items)	0.24	**0.56**	**0.33**	0.27
French	1 (logistic regression, all items)	**1.00**	0.13	0.22	**0.15**
	2 (adjusted, longer items only)	0.50	0.13	0.20	0.00
	3 (adjusted, all items)	0.33	**0.18**	**0.23**	0.12

Table 2: Results

References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. Feel: a French expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855.

Elise Bigeard, Natalia Grabar, and Frantz Thiessard. 2018. Detection and analysis of drug misuses. a study based on social media messages. *Frontiers in pharmacology*, 9:791.

Tilia Ellendorff, Joseph Cornelius, Heath Gordon, Nicola Colic, and Fabio Rinaldi. 2018. UZH@SMM4H: System descriptions. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 56–60.

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen O'Connor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. 2014. Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*, pages 1–8.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 595.

Serge Heiden, Jean-Philippe Magué, and Bénédicte Pincemin. 2010. TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Sergio Bolasco, Isabella Chiari, and Luca Giuliano, editors, *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, volume 2, pages 1021–1032, Rome, Italy. Edizioni Universitarie di Lettere Economia Diritto.

Ferdaous Jenhani, Mohamed Salah Gouider, and Lamjed Ben Said. 2016. A hybrid approach for drug abuse events extraction from Twitter. *Procedia computer science*, 96:1032–1040.

Ferdaous Jenhani, Mohamed Salah Gouider, and Lamjed Ben Said. 2019. Hybrid system for information extraction from social media text: Drug abuse case study. *Procedia Computer Science*, 159:688–697.

Svetlana Kiritchenko, Saif M Mohammad, Jason Morin, and Berry de Bruijn. 2017. NRC-Canada at SMM4H shared task: Classifying tweets mentioning adverse drug reactions and medication intake. In *SMM4H@ AMIA*.

Ari Z. Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (#smm4h) shared tasks at coling 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*.

Alex Lamb, Michael Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795.

Anne-Lyse Minard, Christian Raymond, and Vincent Claveau. 2018. Participation de l'IRISA à DeFT 2018 : classification et annotation d'opinion dans des tweets. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, page 265.

Aurélie Névéol, Julien Grosjean, Stéfan Jacques Darmoni, Pierre Zweigenbaum, et al. 2014. Language resources for French in the biomedical domain. In *Proceedings of LREC*, pages 2146–2151.

Karen O'Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on Twitter? Mining tweets for adverse drug reactions. In *AMIA annual symposium proceedings*, volume 2014, page 924. American Medical Informatics Association.

Abeed Sarker and Graciela Gonzalez-Hernandez. 2017. Overview of the second social media mining for health (SMM4) shared tasks at amia 2017. *Training*, 1(10,822):1239.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez. 2016. Social media mining shared task workshop. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 581–592.

Assaf Urieli. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Phd thesis, University of Toulouse, France.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. 2018. Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 13–16.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. 2019. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 21–30.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.