Automatic Detection of Stigmatizing Uses of Psychiatric Terms on Twitter

Véronique Moriceau, Farah Benamara, Abdelmoumene Boumadane

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 237–243
Marseille, 20-25 June 2022

Automatic Detection of Stigmatizing Uses of Psychiatric Terms on Twitter

Véronique Moriceau, Farah Benamara, Abdelmoumene Boumadane
IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
{firstname.lastname}@irit.fr

Abstract

Psychiatry and people suffering from mental disorders have often been given a pejorative label that induces social rejection. Many studies have addressed discourse content about psychiatry on social media, suggesting that they convey stigmatizing representations of mental health disorders. In this paper, we focus for the first time on the use of psychiatric terms in tweets in French. We first describe the annotated dataset that we use. Then we propose several deep learning models to detect automatically (1) the different types of use of psychiatric terms (medical use, misuse or irrelevant use), and (2) the polarity of the tweet. We show that polarity detection can be improved when done in a multitask framework in combination with type of use detection. This confirms the observations made manually on several datasets, namely that the polarity of a tweet is correlated to the type of term use (misuses are mostly negative whereas medical uses are neutral). The results are interesting for both tasks and it allows to consider the possibility for performant automatic approaches in order to conduct real-time surveys on social media, larger and less expensive than existing manual ones.

Keywords: mental health, opinion analysis, social media

1. Introduction

Mental health stigma finds its roots in the history of psychiatry, in its connection to madness representations. People suffering from mental disorders have often been given a pejorative label that induces social rejection with multiple negative impacts such as difficulties for professional integration, access to housing or interpersonal relationships (Giordana, 2010; Crisp et al., 2000; Lampropoulos et al., 2018). Difficulties concern also the treatment itself, including delay in initial medical consultation, difficulty in accepting the illness, tenuous therapeutic alliance, etc. (Giordana, 2010).

Many studies analyzing newspaper articles pointed out a major diversion from the use of psychiatric terms (Athanasopoulou and Välimäki, 2014; Magliano et al., 2011). A French survey conducted by the OBSoCo (L'Observatoire Société et Consommation) (L'OBSoCo, 2015) noticed that the French terms for *schizophrenic* and *schizophrenia* are particularly used in the context of violent news items and are often employed metaphorically with a negative connotation (e.g. *A policewoman stabbed by a man oscillating between schizophrenia and radicalization*).

With the raise of the internet and social media, it becomes important to analyze how psychiatric terms are employed by general people to act effectively against stigma. Indeed, from the internet users' point of view, Berry et al. (2017) showed that tweeting about mental health helps reducing isolation, fighting stigmatization and raising awareness of mental health by improving knowledge, promoting free expression and strengthening coping and empowerment strategies.

In this paper, we focus for the first time on the use of psychiatric terms in tweets in French. The goal is to analyze if these terms are used in a negative context and/or are misused, assuming that a cause of stigma may be the misuse of psychiatric terms in a negative context (a lot of psychiatric terms being often used as insults). Our contributions are:

- A multilayer annotation scheme that includes the type of term use (medical usage, misuse, irrelevant usage) and the polarity of the tweet (positive, negative, neutral/mixed).
- A dataset of about 22,579 tweets in French containing a wide range of psychiatric terms, in particular nosographic terms relative to psychiatric disorders but also generic and therapeutic terms. A subset of this dataset composed of 3,242 tweets has been manually annotated by clinical psychiatrics. The dataset will be made available to the research community.
- Deep learning models to detect automatically the type of use and the polarity of tweets relying on monotask and multitask architectures. The results show that multitask detection is the most effective and empirically confirm the strong correlation between the type of term use and the polarity of the tweet where these terms appear (misuses are mostly negative whereas medical uses are neutral). Our results are encouraging and constitute an important first step towards mental health stigma detection in French social media.

In the following section, we present an overview of the main studies concerning psychiatric terms on social media. Section 3 describes the dataset, the annotation guidelines and the main observations drawn from the annotation campaign. In Section 4, we present our deep learning models for an automatic classification of tweets according to term use and polarity.

2. Related Work

2.1. Psychiatry on Social Media: Corpus-based Studies

Since 2014, many studies have addressed discourse content about psychiatry on Twitter, suggesting that social networks convey stigmatizing representations of mental health and its users. To our knowledge, existing studies focus on a limited number of psychiatric disorders terms such as *depression*, *schizophrenia* and *autism*, and focus mainly on English, Greek or Chinese. For example, Lachmar et al. (2017) created the hashtag *#MDLL* (*#mydepressionlookslike*) and analyzed 3,225 tweets highlighting seven topics when Twitter users talk about depression: dysfunctional thoughts, impact on daily life, social difficulties, hiding behind a mask, sadness and apathy, suicidal behaviors/ideas, and seeking support/help. Reavley and Pilkington (2014) analyzed a corpus of tweets about schizophrenia and depression in English. They found that 5% of tweets related to schizophrenia convey stigmatizing remarks while less than 1% concern depression. In addition, in their dataset, they found that the polarity is mostly positive (65% of the tweets) when writing about depression while it is rather neutral (43%) for schizophrenia. Joseph et al. (2015) found that tweets containing the hashtag *#schizophrenia* convey a negative sentiment more frequently than tweets containing *#diabetes* (21% vs. 12.6%). Similarly, Athanasopoulou and Sakellari (2016) showed that tweets about schizophrenia in Greek tend to be more negative, medically inappropriate, sarcastic and used in a non-medical way than tweets about diabetes.

Robinson et al. (2017) analyzed and compared messages about five psychiatric disorders (autism, depression, eating disorders, OCD (Obsessive-Compulsive) and schizophrenia) and five physical diseases (AIDS, asthma, cancer, diabetes and epilepsy). In their corpus, schizophrenia and HIV are the most stigmatized diseases and are perceived as dangerous and with an uncontrollable and unpredictable nature. The authors found more than 40% of stigmatizing tweets are about schizophrenia compared to less than 5% of those about depression. Alvarez-Mon et al. (2019) studied the use of the term *psychosis* and compared it to *diabetes, HIV, Alzheimer's disease, breast cancer*. The results show a predominance of non-medical content (33.3%) with a high frequency of misuse and a pejorative opinion tone (36.2%) in the tweets related to psychosis compared to the tweets related to the physical diseases studied.

Finally, Li et al. (2020) recently examined the psycholinguistic characteristics of schizophrenia-related stigma on Sina Weibo, a Chinese microblogging website, and explored whether schizophrenia-related stigma can be distinguished from depression-related stigma in terms of psycholinguistic style. They found that 26.22% of schizophrenia-related posts were labeled as stigmatizing posts and that the proportion of posts indicating depression-related stigma was lower than that indicating schizophrenia-related stigma.

2.2. Automatic Detection of Mental Disorders

Concerning automatic approaches, mental health on social media has been the focus of several shared tasks such as Computational Linguistics and Clinical Psychology (CLPsych)[1] (Goharian et al., 2021) or eRisk[2] (Parapar et al., 2021), dedicated to the (early) detection of users suffering from a specific mental disorder such as depression or eating disorders. Models developed for mental disorders automatic detection on social media are either feature-based (use of personal pronouns, positive or negative emotion, etc. (Bae et al., 2021)) or deep learning models (Wang et al., 2020) (see (Ríssola et al., 2021) for a survey of computational methods for mental state assessment on social media).

Compared to existing corpus-based studies, we propose the first analysis of psychiatric terms in French tweets that goes beyond a small set of nosographic terms. To our knowledge, these studies focusing on the stigma of mental disorders are not automatic, and existing automatic methods deal only with the detection of users suffering from a particular mental disorder (see (Harrigian et al., 2021) for a survey of existing datasets and tasks for mental health research on social media).

3. An Annotated Corpus for Psychiatric Terms

3.1. Data Collection

The main objectives of the study are to analyze (1) how psychiatric terms are used on Twitter, in particular whether they are employed in a medical use or not, and (2) the opinion polarity towards these terms. Our assumption is that psychiatric terms are often misused and that these misuses have a negative polarity.

The corpus is composed of tweets in French that contain at least one of the 90 selected terms[3] relative to psychiatry grouped according to three dimensions:

- **Generic** terms via the stem *psychiatr* which allows to collect morphological variations such as *psychiatrie, psychiatrique, psychiatre* (*psychiatry, psychiatric, psychiatrist*),
- **Nosographic** terms relative to psychiatric disorders. Following the Diagnostic and Statistical Manual of Mental Disorder taxonomy (American Psychiatric Association, 2015), terms are grouped into five categories: Schizophrenia spectrum and other psychotic disorders (e.g. *psycho, schizophrenia*), Bipolar and depressive disorders (e.g. *bipolar, manic*), Autism spectrum disorders (e.g. *austism, autistic*), Anxiety disorders (e.g.

[1] https://clpsych.org/

[2] https://erisk.irlab.org/

[3] The terms have been selected by clinical psychiatrics.

phobic, obsessive-compulsive), and finally Other disorders (e.g. *anorexia, attention deficit hyperactivity disorder*),

- **Therapeutic** terms relative to the most used drugs in the psychiatry field (e.g. *Xanax, alprazolam,* etc.).

After removing retweets and duplicates, a total of 22,579 tweets were collected from 01/01/2016 to 12/31/2018 (see Table 1).

Psychiatric terms	**#tweets**	**#annotated**
Generic	6,993	1,086
Nosographic	12,149	2,604
Schizophrenia spectrum	1,304	1,300
Bipolar/depressive disorders	3,500	647
Autism spectrum	4,389	232
Anxiety disorders	5,855	400
Other disorders	101	25
Therapeutic	1,853	160

Table 1: Number of tweets and annotated tweets containing the selected terms *(a tweet may contain several keywords)*.

3.2. Annotation

A multilayered annotation scheme has been defined that aims at answering two main questions: Are the psychiatric terms in the tweet employed in a medical use or not? What is the overall opinion given in the tweet? We detail below each layer.

3.2.1. Types of term use

The tweet can be annotated according to three possible types of use for psychiatric terms: medical use, misuse or irrelevant use, as follows:

- **Medical use**: it corresponds to the medical definition of the term. The term is used to refer to a medical pathology or to the domain of psychiatry, as in examples (1) and (2).

  (1) *Tellement dégueulasse le valium en gouttes (Oral valium is so disgusting)*

  (2) *Tout à l'heure g écouter une vidéo des voix qu'les schizo entendent dans leurs têtes g pas pu tenir + de 30sec g cru devenir folle (I listened to a video of voices heard by schizophrenic people I coudn't hold more than 30sec I thought I was going insane)*

- **Misuse**: when a psychiatric term is used in a figurative or metaphoric way, as in (3) and (4).

  (3) *Là j'suis en colère tu changes toutes les minutes, à croire que t'es bipolaire. (Now I'm angry you're changing your mind every minute, I'd think you're bipolar)*

  (4) *Tu viens d faire quoi sale autiste (What have you just done, you f*** autistic)*

- **Irrelevant use**: when the tweet is not understandable (lack of context, link towards an URL, advertising, etc.) or not relevant to psychiatry (use of synonyms), as in (5).

  (5) *Les piles au lithium peuvent prendre feu, et les pilotes n'en veulent pas dans les avions (Lithium batteries can catch fire, and pilots don't want them on airplanes)*

3.2.2. Polarity of the tweet

Concerning relevant uses, three standard possible values for tweet polarity are considered: positive, negative or neutral (neutral includes in addition mixed opinion to account for cases where the opinion can be positive and negative at the same time). Opinion polarity is considered at the tweet level regardless if the expressed opinion is related or not to a psychiatric term. Indeed, our aim is to verify if psychiatric terms are used in a positive, negative or neutral context.

- **Positive polarity**: A tweet is annotated as positive when the author expresses a positive personal opinion on facts, events or on a quote; when the tweet is in favor of psychiatry as in (6) or when the author defends the proper medical use of psychiatric terms regardless of their valence as in (7):

  (6) *Mon Rdv psychiatre de demain tombe à la perfection. Pour une fois je l'avoue, j'en ai grandement besoin. (Tomorrow is the perfect timing for my psychiatric appointment. To be honest, for once, I really need it)*

  (7) *Bipolaire c'est un vrai trouble psychiatrique, mesdames arrêtez de le mettre en TN vous n'êtes pas bipolaires vous êtes juste mal éduquées. (Bipolar defines a real mental disorder. Ladies, stop using this term as tweet name. You are not bipolar, you are just poorly-educated)*

- **Negative polarity**: A tweet is annotated as negative when the author expresses a negative personal opinion on facts, events or on a quote as in (8); when the tweet includes insults or ironic/sarcastic comments (cf. (9)) or reports negative facts connected to psychiatry (cf. (10)):

  (8) *La psychiatrie ça brise encore plus les gens. (Psychiatry breaks people down even more)*

  (9) *La France est une terre d'asile... psychiatrique ! (France is a land of asylum... psychiatric asylum!)*

  (10) *Paris : la psychiatre vendait de faux certificats médicaux aux envahisseurs sans-papiers (Paris: a psychiatrist used to sell fake medical certificates to paperless invaders)*

- **Mixed/neutral polarity**: A tweet is annotated as neutral when the opinion of the author is not clearly expressed (cf. (11)) or when it is mixed, both positive and negative (cf. (12)):

  (11) *Lundi j'ai été mise dans la section psychiatrique d'un hôpital. Cette section est pour les personnes entre 10 et 15ans. (On Monday I was put in the psychiatric section of a hospital. This section is for people from 10 to 15 years old)*

  (12) *La psychiatrie c'est cool, Faire ça dans un lieu de stage où ils te harcèlent jusqu'à la dernière heure de tout ton stage par contre moins. (Psychiatry is fun but throughout the internship they badger you, it's less fun)*

Tweets have been manually annotated by two French native speakers, both clinical psychiatrics. They were first trained on 157 tweets and then they annotated separately the same 319 tweets so that an inter-annotator agreement could be computed (Cohen's kappa = 0.829 for type of use and 0.817 for polarity). In the end, 3,242 tweets have been manuallly annotated (see Table1).

3.3. Observations

Table 1 provides the distribution of each type of term in the annotated corpus. Tweets containing diagnostic terms are the most frequent and schizophrenia spectrum terms are dominant. Table 2 shows the distribution of tweets for each class.

Among the 3,242 annotated tweets, 12% are annotated as irrelevant, 45.28% as medical use and 42.72% as misuse. Concerning polarity, 50.37% are annotated as having a negative polarity. Among tweets annotated as misuse, 85.78% are negative while only 0.72% are positive. Furthermore, 19.01% of the tweets annotated as medical use have a positive polarity. It is interesting to note that most tweets annotated as medical use are neutral whereas tweets annotated as misuse are mostly negative. These results confirm the existence of stigmatizing term uses and negative prejudices related to psychiatry and mental disorders as there is a high correlation between the type of use and polarity (using, the χ^2 test, $\chi^2 = 920.04$, $df = 2$, $p < 0.005$). More details can be found in (Delanys et al., 2022).

	medical	misuse	irrelevant
TOTAL	1,468	1,385	389
positive	279	10	
negative	445	1,188	
neutral/mixed	744	187	

Table 2: Distribution for each type of use and polarity.

4. Experiments and Results

4.1. Models

We experimented several feature-based (SVM) and deep learning models (CNN, LSTM, transformers) but we only report here the models having the best results for one or both of our classification tasks, namely type of use of psychiatric terms and tweet polarity detection. A pre-processing step removed URLs and mentions from the tweets and replaced numbers with a tag.

- **BERT$_{\text{base}}$** relies on the pre-trained BERT multilingual cased model (Devlin et al., 2019). We used the HuggingFace's PyTorch implementation of BERT (Wolf et al., 2019) that we trained for four epochs using a gradient clipping of 1.0.
- **FlauBERT$_{\text{base}}$** uses the FlauBERT base cased model (Le et al., 2019), the pre-trained French contextual embeddings. We run the HuggingFace's PyTorch implementation of FlauBERT for four epochs and a learning rate of $2e-5$. For better convergence, we use the linear decreasing learning rate during optimisation. To avoid exploding gradients, we use a gradient clipping of 1.0. This model and the previous one are considered as strong baselines.
- **BERT$_{\text{tuned}}$** and **FlauBERT$_{\text{tuned}}$**: we fine-tuned BERT$_{\text{base}}$ and FlauBERT$_{\text{base}}$ language models initially trained on a general domain by using the 19,337 unlabeled tweets of the dataset. For both models, adaptation consists of training the language models with a masked language model head then use the shifted weights to perform the classification. This process is similar to the initial BERT training.
- **FlauBERT$_{\text{sampling}}$** and **FlauBERT$_{\text{tuned-sampling}}$**: The dataset being relatively small, FlauBERT$_{\text{base}}$ (resp. FlauBERT$_{\text{tuned}}$) is used with a data sampler which does oversampling for low frequent classes and undersampling for high frequent ones when populating each batch[4]. Our aim here is to compare with an effective approach for handling imbalanced data.
- **FlauBERT$_{\text{multitask}}$** and **FlauBERT$_{\text{tuned-multitask}}$**: Following (Liu et al., 2019), the models are used in a multitask learning framework considering there are two classification tasks (type of use and polarity). In FlauBERT$_{\text{multitask}}$ (resp. FlauBERT$_{\text{tuned-multitask}}$), the classifiers for both tasks share and update the same low layers of FlauBERT$_{\text{base}}$ (resp. FlauBERT$_{\text{tuned}}$) except the final task-specific classification layer. We trained the models for 10 epochs with a lr of $2e-5$ using an Adam optimizer (Kingma and Ba, 2015).

[4] https://github.com/ufoym/imbalanced-dataset-sampler

Models	Type of use			Polarity		
	Precision	Recall	F-score	Precision	Recall	F-score
SVM (BoW)	0.774	0.611	0.629	0.686	0.465	0.465
CNN (FastText embeddings)	0.630	0.598	0.610	0.187	0.333	0.239
$\text{BERT}_{\text{base}}$‡	0.7366	0.6926	0.7096	0.5195	0.5056	0.5111
$\text{BERT}_{\text{tuned}}$	0.7240	0.7048	0.7128	0.5536	0.5345	0.5412
$\text{FlauBERT}_{\text{base}}$‡	0.7963	0.7296	0.7531	0.6186	0.6066	0.6096
$\text{FlauBERT}_{\text{base-sampling}}$	0.7813	0.7629	0.7705	0.5885	0.6024	0.5945
$\text{FlauBERT}_{\text{tuned}}$	0.8028	0.7780	0.7886	0.6816	0.6262	0.6452
$\text{FlauBERT}_{\text{tuned-sampling}}$	0.8161	0.7813	**0.7952**	0.6676	0.6419	0.6527
$\text{FlauBERT}_{\text{tuned}}$+F	0.7841	0.7628	0.7714	0.7184	0.6889	0.7011
$\text{FlauBERT}_{\text{multitask}}$	0.7700	0.7327	0.7459	0.6925	0.6315	0.6521
$\text{FlauBERT}_{\text{tuned-multitask}}$	0.7977	0.7393	0.7607	0.7293	0.6767	0.6972
$\text{FlauBERT}_{\text{tuned-multitask}}$+F	0.8085	0.7693	0.7847	0.7268	0.7042	**0.7143**

Table 3: Results for both classification tasks: type of use and polarity in terms of precision, recall and macro-F1 score. ‡: baseline models. +F: with extra-features.

- **$\text{FlauBERT}_{\text{tuned+F}}$** and **$\text{FlauBERT}_{\text{tuned+multitask+F}}$**: Our aim is to test whether additional features can improve over a transformer architecture. Thus, we also experimented with multi-input models that use extra-features added on top of pre-trained contextual word embeddings, among which: tweet meta features (number of likes and the number of retweets of each tweet,[5]) emoji features (number of positive and negative emojis),[6] and opinion features (the averaged number of positive, negative and neutral words in each tweet) relying both on opinion (Benamara et al., 2014) and emotion (Piolat and Bannour, 2009) French lexicons.

4.2. Results and Discussion

Table 3 shows the results obtained for both tasks on the annotated dataset (80% for training, 20% for testing). Even if the dataset is rather small, the results are interesting. We observe that fine-tuned models have better results for both tasks. We also note that adding extra-features on top of the models is very productive for the polarity task, the best two models for this task being $\text{FlauBERT}_{\text{tuned}}$+F and $\text{FlauBERT}_{\text{tuned-multitask}}$+F.

Table 4 shows the results per class for the best model for type of use detection, namely $\text{FlauBERT}_{\text{tuned-sampling}}$. The results are good for both classes *medical use* and *misuse* but are still lower for the minority class (*non relevant*) mainly because of a lower recall, although a sampling method has been applied.

Table 5 shows the results per class for the best model for polarity detection, namely $\text{FlauBERT}_{\text{tuned-multitask}}$+F. Unsurprisingly, results are better for the majority class (*negative*) and are rather equivalent for the other two classes. It is interesting to note that the best results for the polarity task are obtained in a multitask framework where the polarity classifier shares information from the type of use classifier. It confirms the fact that the polarity of a tweet is correlated to the type of use of a psychiatric term in this tweet.

Type of use	Precision	Recall	F-score
non relevant	0.7812	0.6329	0.6993
misuse	0.8566	0.8157	0.8357
medical use	0.8105	0.8953	0.8508

Table 4: Results per class for type of use classification ($\text{FlauBERT}_{\text{tuned-sampling}}$).

Polarity	Precision	Recall	F-score
negative	0.8388	0.8781	0.8580
positive	0.6557	0.5714	0.6107
neutral/mixed	0.6857	0.6630	0.6742

Table 5: Results per class for polarity classification ($\text{FlauBERT}_{\text{tuned-multitask}}$+F).

When analyzing the classification errors, we noticed that tweets containing the stem *psychiatr-* used to collect generic terms represent 16% of the misclassified tweets for the type of use task.

Concerning the polarity task, tweets containing the stem *psychiatr-* represent 38% of the misclassified instances and among them, 47% are misclassified as negative opinion. 45% of the misclassified instances are tweets annotated as *neutral/mixed opinion* and among them, 71.4% have been misclassified as negative opinion.

These observations suggest that, in both tasks, tweets containing generic terms (*psychiatry, psychiatric, ...*) may be more difficult to classify than tweets containing nosographic or therapeutic terms. It also suggests that

[5]We also experimented with the number of followers, and user mentions but the results were lower.

[6]We relied on a manually built emoji lexicon that contains 1,644 emojis along with their polarity and detailed description.

the predictions for the polarity task have probably been biased by the majority class (*negative opinion*) since the best model for this task does not use data sampling.

5. Conclusion

In this paper, we focused for the first time on the stigmatizing uses of psychiatric terms in tweets in French and proposed several deep learning models to detect automatically (1) the different types of use of psychiatric terms (medical use, misuse or irrelevant use), and (2) the polarity of the tweet, assuming that a cause of stigma may be the misuse of psychiatric terms in a negative context. The results show that polarity detection can be improved when done in a multitask framework in combination with type of use detection. This confirms the observations made manually on several other datasets dedicated to mental health: the polarity of a tweet is correlated to the type of term use (misuses of psychiatric terms are mostly negative whereas medical uses are neutral). Event if the dataset used is relatively small, the results are interesting for both tasks and it allows to consider the possibility for performant automatic approaches in order to conduct real-time surveys on social media, larger and less expensive than existing manual ones.

6. Ethical Approval

This article does not contain any studies with human participants carried out by any of the authors. In addition, the data that was used is composed of textual content from the public domain and the dataset conforms to the Twitter Developer Agreement and Policy that allows unlimited distribution of the numeric identification number of each tweet.

However, it is important to note that in this work, we are not interested in detecting mental disorders or social media users suffering from these disorders, but in detecting in which context psychiatric terms related to several mental disorders are used in order to highlight stigmatizing uses.

Moreover if any of the users want to opt out from having their data being used for research, they can request that they be removed from the dataset by sending an email to the authors of this paper.

Acknowledgements

This work has been carried out in the framework of the STERHEOTYPES project funded by the Compagnia San Paolo 'Challenge for Europe'.

7. Bibliographical References

Alvarez-Mon, M., Llavero-Valero, M., Sánchez-Bayona, R., Pereira-Sanchez, V., Vallejo-Valdivielso, M., Monserrat, J., Lahera, G., and Asunsolo del Barco, A. (2019). Areas of Interest and Stigmatic Attitudes of the General Public in Five Relevant Medical Conditions: Thematic and Quantitative Analysis Using Twitter. *Journal of Medical Internet Research*.

Athanasopoulou, C. and Sakellari, E. (2016). "Schizophrenia" on Twitter: Content Analysis of Greek Language Tweets. *Studies in Health Technology and Informatics*, 226.

Athanasopoulou, C. and Välimäki, M. (2014). "Schizophrenia" as a Metaphor in Greek Newspaper Websites. *Studies in Health Technology and Informatics*, 202.

Bae, Y., Shim, M., and Lee, W. (2021). Schizophrenia detection using machine learning approach from social media content. *Sensors (Basel)*, 21(17).

Benamara, F., Moriceau, V., and Mathieu, Y. Y. (2014). Fine-grained semantic categorization of opinion expressions for consensus detection (in French). In *DEFT 2014 Workshop: Text Mining Challenge*, pages 36–44.

Berry, N., Lobban, F., Belousov, M., Emsley, R., Nenadic, G., and Bucci, S. (2017). #WhyWeTweetMH: Understanding Why People Use Twitter to Discuss Mental Health Problems. *Journal of Medical Internet Research*, 19(4).

Crisp, A., Gelder, M., Rix, S., Meltzer, H., and Rowlands, O. (2000). Stigmatisation of people with mental illness. *The British Journal of Psychiatry*, 177(1).

Delanys, S., Benamara, F., Moriceau, V., Mothe, J., and Olivier, F. (2022). Psychiatry on Twitter: A content analysis of the use of psychiatric terms in French. *JMIR Formative Research*, 6(2).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL:HLT, Volume 1 (Long and Short Papers)*.

Giordana, J.-Y. (2010). *La stigmatisation en psychiatrie et en santé mentale*. Elsevier Masson.

N. Goharian, et al., editors. (2021). *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*.

Harrigian, K., Aguirre, C., and Dredze, M. (2021). On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*.

Joseph, A., Tandon, N., Yang, L., Duckworth, K., Torous, J., Seidman, L., and Keshavan, M. (2015). #Schizophrenia: Use and misuse on Twitter. *Schizophrenia Research*, 165(2–3).

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*.

Lachmar, E., Wittenborn, A., Bogen, K., and McCauley, H. (2017). #MyDepressionLooksLike: Examining Public Discourse About Depression on Twitter. *JMIR Mental Health*.

Lampropoulos, D., Fonte, D., and Apostolidis, T. (2018). La stigmatisation sociale des personnes vi-

vant avec la schizophrénie : une revue systématique de la littérature. *L'évolution psychiatrique*, 43(1).

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2019). FlauBERT: Unsupervised Language Model Pre-training for French. *arXiv preprint arXiv:1912.05372*.

Li, A., Jiao, D., Liu, X., and Zhu, T. (2020). A comparison of the psycholinguistic styles of schizophrenia-related stigma and depression-related stigma on social media: Content analysis. *J Med Internet Res.*, 22(4).

Liu, X., He, P., Chen, W., and Gao, J. (2019). Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the ACL, Volume 1: Long Papers*.

L'OBSoCo. (2015). L'image de la schizophrénie à travers son traitement médiatique.

Magliano, L., Read, J., and Marassi, R. (2011). Metaphoric and non-metaphoric use of the term "schizophrenia" in Italian newspapers. *Social Psychiatry and Psychiatric Epidemiology*, 46(10).

Parapar, J., Martín-Rodilla, P., Losada, D., and Crestani, F. (2021). Overview of erisk at CLEF 2021: Early risk prediction on the internet (extended overview). In *Conference and Labs of the Evaluation Forum (CLEF)*.

Piolat, A. and Bannour, R. (2009). An example of text analysis software (EMOTAIX-Tropes) use: The influence of anxiety on expressive writing. *Current psychology letters*, 25.

Reavley, N. and Pilkington, P. (2014). Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. *PeerJ*, 2:e647.

Ríssola, E., Losada, D., and Crestani, F. (2021). A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare*, 2(2).

Robinson, J., Bailey, E., Hetrick, S., Paix, S., O'Donnell, M., Cox, G., Ftanou, M., and Skehan, J. (2017). Developing Social Media-Based Suicide Prevention Messages in Partnership With Young People: Exploratory Study. *JMIR Mental Health*, 4(4).

Wang, Y., Wang, Z., Li, C., Zhang, Y., and Wang, H. (2020). A Multitask Deep Learning Approach for User Depression Detection on Sina Weibo. arXiv preprint arXiv:2008.11708.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

8. Language Resource References

American Psychiatric Association. (2015). *DSM-5: Manuel diagnostique et statistique des troubles mentaux*. Elsevier Masson.