Between the Lines: Machine Learning for Prediction of Psychological Traits - A Survey

Dirk Johannßen, Chris Biemann

Between the Lines: Machine Learning for Prediction of Psychological Traits - a Survey

Dirk Johannßen and Chris Biemann

LT Group, MIN Faculty, Dept. of Computer Science, Universität Hamburg,
Vogt-Kölln-Straße 30, 22527 Hamburg, Germany
{johannssen,biemann}@informatik.uni-hamburg.de
http://lt.informatik.uni-hamburg.de/

Abstract. A connection between language and psychology of natural language processing for predicting psychological traits (NLPsych) is apparent and holds great potential for accessing the psyche, understand cognitive processes and detect mental health conditions. However, results of works in this field that we call NLPsych could be further improved and is sparse and fragmented, even though approaches and findings often are alike. This survey collects such research and summarizes approaches, data sources, utilized tools and methods, as well as findings. Approaches of included work can roughly be divided into two main strands: word-list-based inquiries and data-driven research. Some findings show that the change of language can indicate the course of mental health diseases, subsequent academic success can be predicted by the use of function words and dream narratives show highly complex cognitive processes – to name but a few. By surveying results of included work, we draw the 'bigger picture' that in order to grasp someone's psyche, it is more important to research *how* people express themselves rather than *what* they say, which surfaces in function words. Furthermore, often research unawarely induce biases that worsen results, thus leading to the conclusion that future research should rather focus on data-driven approaches rather than hand-crafted attempts.

Keywords: computational psychology · machine learning · survey · natural language processing

1 Introduction

One rather newly opened application field for natural language processing (NLP), is NLP for predicting psychological traits, which we call NLPsych. Due to computer systems in clinical psychology, massive amounts of textual interactions in social networks, as well as an uprising of blogs and online communities, the availability of massive amounts of data has catalyzed research of psychological phenomena such as mental diseases, connections between intelligence and use of language or a data-driven understanding of dream language – to name but a few – with NLP methods.

Possible applications range from detecting and monitoring a course of mental health illnesses by analyzing language [1], finding more objective measures and language clues on subsequent academic success of college applicants [2] or discovering that dream narratives show highly complex cognitive processes [3]. Promising possible scenarios for future work could explore connections of personality traits or characteristics with subsequent development or research on current emotional landscapes of people by their use of language.

Even though the potential is high, the sub-field of natural language processing in psychology we call NLPsych henceforth is a rather fragmented field. Results of included works vary in accuracy and often show room for improvement by using either best-practice methods, by shifting the research focus onto e.g. function words, by using data-driven methods or by combining established approaches in order to perform better. This survey provides an overview over some of the recent approaches, utilized data sources and methods, as well as findings and promising pointers. Furthermore, the aim is to hypothesize about possible connections of different findings in order to draw a 'big picture' from those findings.

1.1 Research questions

Even though the most broadly employed research questions target mental health due to the high relevance of findings in clinical psychology, this survey ought to have a broader understanding. Thus the following research questions, which have been derived from included work, approach mental changes, cognitive performance and emotions. Three exemplary works, that address similar questions, are mentioned as well. Important ethical considerations are out of scope of this paper (see e.g. dedicated wokrshops[1]).

Research question i) Does a change of the cognitive apparatus also change the use of language and if so, in what way? (e.g. [4][3][5])

Research question ii) Does the use of language correspond to cognitive performance and if so, which aspects of language are indicators? (e.g. [6][7][2])

Research question iii) Is a current mood or emotion detectable by the use of language besides explicit descriptions of the current mental state? (e.g. [7][1][8])

1.2 Structure of this paper

Firstly, this survey aims to grant an overview of some popular problem domains with an idea of employed approaches and data sources in Section 2 and the development of broadly utilized tools in Section 3. Widely employed measurements of different categories will be discussed (Section 4). Section 5 describes utilized methods and tools, and is divided into two strands: on the one hand, data-driven approaches that use supervised machine learning and on the other hand, statistics on statistics on manually defined features and so-called 'inquiries', i.e.

[1] `http://www.ethicsinnlp.org/`

counts over word lists and linguistic properties. Secondly, this work surveys parallels and connections of important findings that are utilized in order to conclude a 'bigger picture' in Section 6.

1.3 Target audience

This paper targets an audience that is both familiar with basics of natural language processing and psychology, with some experience in machine learning (ML), Deep Learning (DL) or the use of standard tools for those fields, even though some explanations will be provided.

1.4 Criteria for inclusion

Criteria for the inclusion of surveyed work can be divided into three aspects: Firstly, often cited work and very influential findings were included. Secondly, the origin of included work such as well established associations, authors or journals. And lastly, the soundness of the content with this survey's focus in terms of methodology or topic. Only if a work suits at least one – if not all – of those aspects, said work has been included in this survey. E.g. work published by the Association for Computational Linguistics (ACL)[2], which targets NLP, was considered. Well established journals of different scientific fields such as e.g. Nature, which dedicates itself to natural science, were considered. Search queries included 'lexical database', 'psychometric', 'dream language', 'psychology', 'mental', 'cognitive' or 'text'. Soundness was considered in terms of topic (e.g. subsequent academic success, mental health prediction or dream language), as well as innovative and novel approaches.

2 Popular problem domains, approaches & data sources

This section presents popular NLPsych problem domains and is ordered by descending popularity. Approaches will be briefly explained.

2.1 Mental health

Mental health is the most common problem domain for approaches that use NLP to characterize psychological traits as some of the following works demonstrate.

Depression detection systems. Morales *et al.* [9] summarized different depression detection systems in their survey and show an emerging field of research that has matured. Those depression detection systems often are linked to language and therefore have experienced gaining popularity among NLP in clinical psychology. Morales *et al.* [9] described and analyzed utilized data sources as well. *The Distress Analysis Interview Corpus (DAIC)*[3] offers audio and video

[2] `https://www.aclweb.org/portal/`
[3] `http://dcapswoz.ict.usc.edu/`

recordings of clinical interviews along with written transcripts on depressions and thus is less suitable for textual approaches that solemnly focus on textual data but can be promising when visual and speech processing are included. The *DementiaBank* database offers different multi media entries on the topic of clinical dementia research from 1983 to 1988. *The ReachOut Triage Shared Task* dataset from the SemEval 2004 Task 7 consists of more than 64,000 written forum posts and was fully labeled for containing signs of depression. Lastly, *Crisis Text Line*[4] is a support service, which can be freely used by mentally troubled individuals in order to correspond textually with professionally trained counselors. The collected and anonymized data can be utilized for research.

Suicide attempts. In their more recent work, Coppersmith *et al.* [10] investigated mental health indirectly by analyzing social media behavior prior to suicide attempts on Twitter. *Twitter*[5] is a social network, news- and micro blogging service and allows registered users to post so-called tweets, which were allowed to be 140 characters in length before November 2017 and 280 characters after said date. As before in [11], the Twitter users under observation had publicly self-reported their condition or attempt.

Crisis. Besides depression, anxiety or suicide attempts, there are more general crises as well, which Kshirsagar *et al.* [12] detect and attempt to explain. For their work they used a specialized social network named Koko and used a combination of neural and non-neural techniques in order to build classification models. *Koko*[6] is an anonymous emotional peer-to-peer support network, used by Kshirsagar *et al.* [12]. The dataset originated from a clinical study at the MIT and can be implemented as chatbot service. It offers 106,000 labeled posts, with and some without crisis. A test set of 1,242 posts included 200 crisis labeled entries, i.e. $\sim 16\%$.

Reddit[7] is a community for social news rather than plain text posts and offers many so-called sub-reddits, which are sub-forums dedicated to certain, well defined topics. Those sub-reddits allow for researchers to purposefully collect data. Shen *et al.* [13] detected anxiety on Reddit by using depression lexicons for their research and training Support Vector Machine (SVM, Cortes *et al.* [14]) classifiers, as well as Latent Dirichlet Allocation (LDA, Blei *et al.* [15]) for topic modeling (for LDA see Section 3). Those lexicons offer broad terms that can be combined with e.g. Language Inquirie and Word Count (LIWC, Pennebaker *et al.* [16]) features in order to identify different conditions in order to be able to distinguish those mental health issues. Shen *et al.* [13] used an API offered by Reddit in order to access sub-reddits such as r/anxiety or r/panicparty.

Dementia. In their recent work, Masrani *et al.* [17] used six different blogs to detect dementia by using different classification approaches. Especially the lexical diversity of language was the most promising feature, among others.

[4] `https://www.crisistextline.org/`
[5] `https://twitter.com/`
[6] `https://itskoko.com/`
[7] `https://www.reddit.com/`

Multiple mental health conditions. Coppersmith *et al.* [11] researched the detection of a broad range of mental health conditions on Twitter. Coppersmith *et al.* [11] targeted the well discriminability of language characteristics of the following conditions: attention deficit hyperactivity disorder (ADHD), anxiety, bipolar disorder, borderline syndrome, depression, eating disorders, obsessive-compulsive disorder (OCD), post traumatic stress disorder (PTSD), schizophrenia and seasonal affective disorder (SAD) – all of which were self-reported by Twitter users.

2.2 Dreams language in dream narratives

Dream language. Niederhoffer *et al.* [3] researched the general language of dreams from a data-driven perspective. Their main targets are linguistic styles, differences between waking narratives and dream narratives, as well as the emotional content of dreams. In order to achieve this, they used a community named DreamsCloud. *DreamsCloud*[8] is a social network community dedicated to sharing dreams in a narrative way, which also offers the use of data for research purposes. There are social functions such as 'liking' a dream narrative or commenting on it, as Niederhoffer *et al.* [3] describe in their work. There are more than 119,000 dream narratives from 74,000 users, which makes this network one of the largest of its kind. Since DreamsCloud is highly specialized, issues such as relevance or authenticity are less crucial as they would be on social networks like Facebook[9].

LIWC and personality traits. Hawkins *et al.* [18] layed their focus on LIWC characteristics especially and a correlation with the personality of a dreamer. Data was collected by clinical studies in which Hawkins *et al.* [18] gathered dream reports from voluntary participants. Their work is more thorough in terms of length, depth and rate of conducted experiments on LIWC features. Dreams could be distinguished from waking narratives, but – as of said study – correlations with personality traits could not be found.

2.3 Mental changes

As we will be showing in Section 6, mental changes and mental health problems are seemingly connected. However, natural changes such as growth or life-changing experiences can alter the use of language as well.

Data generation and life-changing events. Oak *et al.* [19] pointed out that the availability of data in the clinical psychology often is a difficulty for researchers. The application scenario chosen for a study on data generation for clinical psychology are life-changing events. Oak *et al.* [19] aimed to use NLP for tweet generation. The BLEU score measures n-gram precision, which can be important for next character- or next word predictions, as well as for classification tasks. Another use case of this measure is the quality of machine translations.

[8] `https://www.dreamscloud.com/`
[9] `http://www.facebook.com`

Oak *et al.* [19] use the BLEU score to evaluate the quality of their n-grams for language production of their data generation approach of life-changing events. Even though the generated data would not be appropriate to be used for e.g. classification tasks, Oak *et al.* [19] nonetheless proposed useful application scenarios such as virtual group therapies. 43 percent of human annotators thought the generated data to be written by real Twitter users.

Changing language over the course of mental illnesses. A study by Reece *et al.* [1] revealed that language can be a key for detecting and monitoring the whole process from onsetting mental illnesses to a peak and a decline as therapy shows positive effects on patients. participants involved in the study had to prove their medical diagnosis and supply their Twitter history. Different techniques were used to survey language changes. MTurk was used for labeling their data. Reece *et al.* [1] were able to show a correlation between language changes and the course of a mental disease. Furthermore, their model achieved high accuracy in classifying mental diseases throughout the course of illness.

Language decline through dementia and Alzheimer's. It is known that cognitive capabilities decline during the course of the illness dementia. Masrani *et al.* [17] were able to show that language declines as well. Lancashire *et al.* [20] researched the possibility of approaching Alzheimer's of the writer Agatha Christie by analyzing novels written at different life stages from age 34 to 82. The first 50,000 words of included novels were inquired with a tool named TACT, which operates comparable to LIWC (shown in Section 3) and showed a decline in language complexity and diversity. During their research, Masrani *et al.* [17] detected dementia by including blogs from medically diagnosed bloggers with and without dementia. Self reported mental conditions, as it is often used for research of social networks, are at risk of being incorrect (e.g. pranks, exaggeration or inexperience).

Development. Goodman *et al.* [8] showed that the acquisition and comprehension of words and lexical categories during the process of growth correspond with frequencies of parental usage, depending on the age of a child. Whilst the acquisition of lexical categories and comprehension of words correlates with the frequency of word usage of parents later on in life, simple nouns are acquired earlier. Thus, whether words were more comprehensible was dependent on known categories and a matter of similarity by the children.

2.4 Motivation and emotion

Emotions and motivations are less common problem domains. Some approaches aim to detect general emotions, further researchers focus on strong emotions such as hate speech, others try to provide valuable resources or access to data.

Distant emotion detection. In order to better understand the emotionality of written content, Pool *et al.* [21] used emotional reactions of Facebook users as labels for classification. *Facebook* offers insightful social measurements such as richer reactions on posts (called *emoticons*) or numbers as friends, even though most available data is rather general.

Hate speech. Serrà *et al.* [4] approached the question of emotional social network posts by surveying the characteristics of hate speech. In order to tackle hate speech usually containing a lot of neologism, spelling mistakes and out-of-vocabulary words (OOV), Serrà *et al.* [4] constructed a two-tier classification that firstly predicts next characters and secondly measures distances between expectation and reality. Other works on hate speech include [22][23][24].

Motivational dataset. Since data sources for some sub-domains such as motivation are sparse, Pérez-Rosas *et al.* [25] created a novel contributing a motivational interviewing (MI) dataset by including 22,719 utterances from 227 distinct sessions, conducted by 10 counselors. *Amazon mechanical turk* (MTurk) is a crowdsourcing service. Research can define manual tasks and define quality criteria. Pérez-Rosas *et al.* [25] used MTurk for labeling their short texts by crowdsourcers. They achieved a high Intraclass Correlation Coefficient (ICC) of up to 0.95. MI is a technique in which the topic 'change' is the main object of study. Thus, as described in Subsection 6.3, this dataset could also contribute to early mental disease detection. MI is mainly used for treating drug abuse, behavioral issues, anxiety or depressions.

Emotions. Pool *et al.* [21] summarized in their section on emotional datasets some highly specialized databases on emotions, which the authors analyzed thoroughly. *The International Survey on Emotion Antecedents and Reactions* (ISEAR)[10] dataset offers 7,665 labeled sentences from 3,000 respondents on the emotions of joy, fear, anger, sadness, disgust, shame and guilt. Different cultural backgrounds are included. *The Fairy Tales*[11] dataset includes the emotional categories angry, disgusted, fearful, happy, sad, surprised and has 1,000 sentences from fairy tales as the data basis. Since fairy tales usually are written with the intention to trigger certain emotions of readers or listeners, this dataset promises potential for researchers. *The Affective Text*[12] dataset covers news sites such as Google news, NYT, BBC, CNN and was composed for the SemEval 2007 Task 14. It offers a database with 250 annotated headlines on emotions including anger, disgust, fear, joy, sadness and surprise.

2.5 Academic success

Few researchers in NLPsych have approached a connection between language and academic success. Some challenges are lack of data and heavy biases as some might assume that an eloquent vocabulary, few spelling mistakes or a sophisticated use of grammar indicate a cognitive skilled writer. Pennebaker *et al.* [2] approached the subject in a data-driven fashion and therefore less biased. Data was collected by accessing more than 50,000 admission essays from more than 25,000 applicants. The college admission essays could be labeled with later

[10] `http://emotion-research.net/toolbox/toolboxdatabase.2006-10-13.2581092615`

[11] `https://github.com/bogdanneacsa/tts-master/tree/master/fairytales`

[12] `http://web.eecs.umich.edu/~mihalcea/downloads/AffectiveText.Semeval.2007.tar.gz`

academic success indicators such as grades. The study showed that rather small words such as function words correlate with subsequent success, even across different majors and fields of study. Function words (also called closed class words) are e.g. pronouns, conjunctions or auxiliary words, which tendentially are not open for expansion, whilst open class words such as e.g. nouns can be added during productive language evolution.

3 Tools

In this section we discuss some broadly used tools for accessing mainly written psychological data. Included frameworks are limited to the programming language Python, since it is well established – especially for scientific computing – and included works mostly use libraries and frameworks designed for Python.

3.1 Word lists

LIWC. The Language Inquiry and Word Count (LIWC) was developed by Pennebaker *et al.* [16] for the English language and has been transferred to other language such as e.g. German by Wolf *et al.* [26]. The tool was psychometrically validated and can be considered a standard in the field. LIWC stands for a tool that operates with recorded dictionaries of word lists and a vector of approximately 96 metrics (depending on the version and language) such as number of pronouns or number of words associated with familiarity to be counted in input texts.

CELEX[13] is a lexical inquiry database, that was developed by Baayen *et al.* [27] and later on enhanced to a CELEX release 2. The database contains 52,446 lemmas and 160,594 word forms in English and a number of those in Dutch and German as well. It is regularly used by researchers such as Fine *et al.* [28] did in order to research possible induced biases in corpora, which used CELEX for predicting human language by measuring proportions of written and spoken English based on CELEX entries.

Kshirsagar *et al.* [12] used the Affective Norms for English Words (ANEW), which is an inquiry tool such as LIWC, as well as labMT, used by Reece *et al.* [1], which is a word list score for sentiment analysis.

3.2 Corpus-induced Models

LDA Blei *et al.* [15] developed a broadly used generative probabilistic model called Latent Dirichlet Allocation (LDA) that is able to collect text through a three-layered Bayesian model that builds models on the basis of underlaying topics.

SRILM. The SRI Language modeling toolkit (SRILM), produced by Stolcke [29] is a software package that consists of C++ libraries, other programs and

[13] `https://catalog.ldc.upenn.edu/ldc96l14`

scripts that combine functionality for processing, as well as producing mainly speech recognition and other applications such as text. Oak *et al.* [19] used SRILM for 4-gram modeling for language generation of life-changing events.

3.3 Frameworks

NLTK. The Natural Language Toolkit (NLTK) is a library for Python that offers functionality for language processing, e.g. tokenization or part-of-speech (POS) tagging. It is used on a general basis. E.g. Shen *et al.* [13] use NLTK for POS tagging and collocation.

Scikit-learn. The tool of choice of Pool *et al.* [21] and Shen *et al.* [13] was scikit-learn [30], a freely available and open sourced library for Python. Since scikit-learn is designed to be compatible with other numerical libraries, it can be considered one of the main libraries for machine learning in the field of natural language processing.

3.4 Further tools

Further tools that are being used in included work in some places are the cross-linguistic lexical norms database (CLEX) [31] for evaluating and comparing early child language, the Berlin affective word list reloaded (BAWL-R) [32] that is based on the previous version of BAWL for researching affective words in the German language and lastly HMMlearn, used by Reece *et al.* [1], which is a Python library for Hidden Markov Models (HMM).

4 Psychometric measures

When conducting research on NLPsych, the selection of psychometric measurements are crucial for evaluating given data on psychological effects or to detecting the presence of target conditions before a classification task can be set up. Therefore, in this section we describe broadly used psychometric measurements of included work. Psychometrics can be understood as a discipline of psychology – usually found in clinical psychology – that focus on 'testing and measuring mental and psychology ability, efficiency potentials and functions' [33].

There are measures for machine learning as well such as e.g. the accuracy score, which will not be covered due to broadly available standard literature on this matter.

4.1 Questionnaires

BDI and HAM-D. The Beck Depression Inventory (BDI) and HAM-D are used and described by Morales *et al.* [9] and Reece *et al.* [1] for measuring the severity of depressions. The HAM-D is a questionnaire that is clinically administrated and consists of 21 questions, whilst the BDI is a questionnaire that consists of the same 21 questions, but is self-reported.

CES-D. The Center for Epidemiological Studies Depression Scale (CES-D) is a questionnaire for participants to keep track on their depression level and has been used in their work by Reece *et al.* [1].

4.2 Wordlist measurements

MITI. The Motivational Interviewing Integrity Treatment score (MITI) measures how well or poorly a clinician is using MI (motivational interviewing), as Pérez-Rosas *et al.* [25] described in their work. The Processes related to 'change talk', thus the topical focus, is the crucial part of this measurement. Global counts and behavior counts distinguish the impact on this measure. Words that encode the MITI level are e.g. 'focus', 'change', 'planning' or 'engagement'.

CDI. The Categorical Dynamic Index (CDI), used by Niederhoffer *et al.* [3], Jørgensen *et al.* [31], as well as Pennebaker *et al.* [2], is described as a bipolar continuum, applicable on any text, that measures the extend of how categorical or dynamic thinking is. Since those two dimensions are said to distinguish between cognitive styles of thinking, it therefore can reveal e.g. whether or not dreamers are the main character of their own dream [3]. The CDI can be measured by inquiring language with tools such as e.g. LIWC and weighting the categories.

5 Broadly used research methods

Since there are two main approaches for performing NLPsych – data-driven approaches and manual approaches from clinical psychology – this section will be divided into those two strands, beginning with feature approaches and ending with data-driven machine learning approaches. Within those strands, the methods are ordered by their complexity.

5.1 A general setup of NLPsych

Even though there are detailed differences between approaches of included works, there is a basic schema in the way NLPsych is set up. Figure 1 illustrates a classification setup. Firstly, after having collected data, pieces of information are read and function as input. Different measures or techniques can be applied to the data by an annotator to assign labels to the input. Whether or not annotation takes place, depends on the task and origin of the data.

Secondly, after separating training, test, and sometimes development sets, features get extracted from those data items, e.g. LIWC category counts, the ANEW sadness score or POS tags. A feature extractor computes a nominal or numerical feature vector, which will be described in Subsection 5.3.

Thirdly, depending on the approach, this feature vector is directly used in rule based models such as e.g. defined LIWC scores that correlate with dream aspects, as Niederhoffer *et al.* [3] did. A different approach uses the feature vector on a machine learning algorithm in order to compute a classifier model,

that thereafter can be used to classify new instances of information, as Reece *et al.* [1] demonstrated in their work.

Finally, for both of the approaches, the accuracy of the classification task is determined and researchers analyze and discuss the consequences of their findings, as well as use the models for classification tasks.

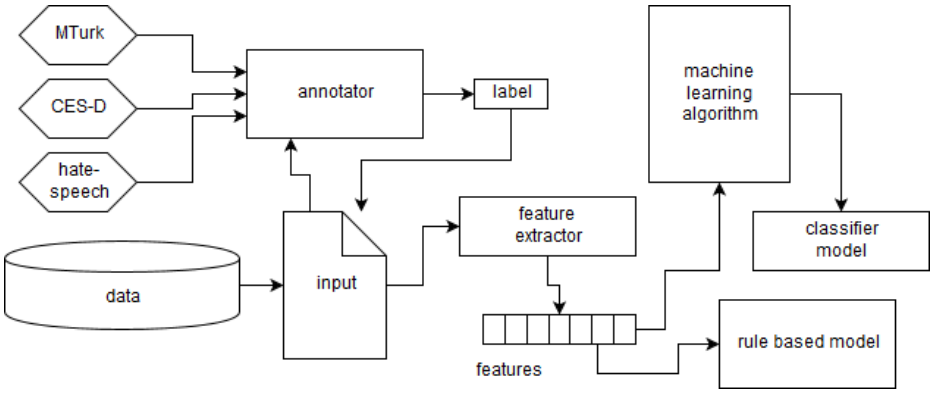

Fig. 1. A general setup for classification tasks in NLPsych

5.2 Supervised machine learning approaches

SVM. Support Vector Machines (SVM) are a type of machine learning algorithm that measure distances of instances to so-called support vectors that map said examples in order to form a dividing gap. This gap separates said examples into categories to perform classification or regression tasks. This broadly utilized standard method has been used by e.g. Pool *et al.* [21] for BOW models via scikit-learn.

HMM. Reece *et al.* [1] used Hidden Markov Models (HMM), which are probabilistic models for modeling unseen events, as well as word shift graphs that visualize changes in the use of language [1].

RNN. The Recurrent Neural Networks (RNN) are an architecture of deep neural networks that differ from feed forward neural networks by having time-delayed connections to cells of the same layer and thus possesses a so-called memory. RNNs require for the input to be numeric feature vectors. Words or sentences typically get transformed by the use of embedding methods (e.g. [34]) into numerical representations. Some authors that use RNNs are Cho *et al.* [35] who used encoder and decoder in order to maximize the conditional probabilities of representations. Kshirsagar *et al.* [12] used RNNs for word embeddings and Serrà *et al.* [4] trained character based language models with RNNs.

LSTM. A Long short-term memory neural network (LSTM, Hochreiter *et al.* [36]) is a type of RNN in which three gates (input, forget and output) in

an inner, so-called memory cell, are employed to be able to learn the amount of retained memory depending on the input and the inner state. LSTMs are capable of saving information over arbitrary steps, thus enabling them to *remember* a short past for sophisticated reasoning. LSTMs nowadays are the method of choice for classification on sequences and can be considered as established standard. Long short-term memory neural networks often are used when calculation power, as well as big amounts of data are available and a memory is needed to train precise models. The latter often is the case when working with psychological data. E.g. Oak *et al.* [19] used an LSTM for training language models for language production of life-changing events.

5.3 Features for characterizing text

Features serve as characteristics of texts and are always computable for every text, e.g. the average rate of words per sentence. Some of said features are numerical, some are nominal. Those features usually are stored in a feature vector that serves as input for classifiers but can be used directly, e.g. in order to perform statistics on them and to draw conclusions. Not every presented feature is being used as such. On the one hand, LIWC, tagging and BOWs are used as characteristics of text and thus are classically used as features. On the other hand, LDA targets the data collection process and n-grams, CLMs, as well as next character predictions can be utilized for modeling.

LIWC. In Section 3 the LIWC is described as a set of categories for which word lists were collected. The core dictionary and tool with its capability of calculating a feature vector for language modeling is well established and can be categorized as method of choice in psychological language inquiry. The way LIWC is used, is very common. However, researchers usually focus on some selected aspects of the feature vector in order to grasp psychological effects. Coppersmith *et al.* [11] used LIWC for differentiating the use of language of healthy people versus people with mental conditions and diseases. Hawkins *et al.* [18] and Niederhoffer *et al.* [3] researched the language landscape of dream narratives. Scores, such as the LIWC sadness score were the basis of the work of Homan *et al.* [5] on depression symptoms. Morales *et al.* [9] also surveyed the broad use of LIWC in depression detection systems. Pennebaker *et al.* [2], which partly developed LIWC used the tool to research word usage in connection with college admission essays. Reece *et al.* [1] captured the general mood of participants by using LIWC and Shen *et al.* [13] surveyed the language of a crisis with LIWC.

LDA. Latent Dirichlet Allocation (LDA) is a probabilistic model for collecting text corpora on the basis of underlaying topics in a three layered bayesian model, as described in Section 3. Some researchers that used the LDA are Niederhoffer *et al.* [3] for topic modeling in order to explore the main themes of given texts and Shen *et al.* [13] which used LDA to predict membership of classes by a given topic.

BOW. Bag of words (BOW – sometimes called vector space models) are models that intentionally dismiss information of the order of text segments or

tokens and thus e.g. grammar by only taking into account presence resp. absence of word types in a text. Usually, BOW models are used for document representation where neither the order nor grammar of tokens are crucial but rather their frequency. Shen *et al.* [13] use so-called continuous bag-of-word models (CBOW, [34]) with a window size of 5 in order to create word embeddings. Homan *et al.* [5], Kshirsagar *et al.* [12] and Serrà *et al.* [4] use BOW for embedding purposes. *Tf-idf* is a measure for relevance that quantifies the term frequency (tf) inverse document frequencies (idf) by using said BOW models [12].

Part of speech tagging (POS). POS is the approach to assign lexical information to segmented or tokenized parts of a text. Those tags can be used as labels and hence be used as additional information for e.g. classification tasks. Some authors that used tagging were Masrani *et al.* [17] and Reece *et al.* [1].

N-grams. A continuous sequence of n tokens of a text is called n-gram. The higher the chosen n, the more precise language models on the basis of n-grams can be used for e.g. classification or language production while training becomes more excessive with higher n. Some of the authors that use either word-based n-grams or character based n-grams are Kshirsagar *et al.* [12], Homan *et al.* [5], Oak *et al.* [19], Reece *et al.* [1] and Shen *et al.* [13].

CLM. A Character n-gram Language Model (CLM) is closely related to n-grams and is a term for language models that use n-gram frequencies of letters for probabilistic modeling, used by Coppersmith *et al.* [11] as model that models emotions on the basis of character sequences.

Next character prediction is the prediction of words of characters on the basis of probabilistic language models, which have been used by Serrà *et al.* [4] for determining the soundness of an expectable use of language with actual language usage in order to detect hate-speech.

Table 1. Overview of included works.

1st Author	Problem	Data sources	Tools	Measures	Method
Morales[9]	Depression	Multiple	i.a. LIWC	BDI, HAM-D	Manual
Copper.[10]	Suicide	Twitter			Manual
Copper.[11]	Multiple	Twitter	LIWC		CLM
Kshirs.[12]	Crisis	Koko	ANEW		RNN
Shen[13]	Anxiety	Reddit	NLTK, LIWC		SVM
Masrani[17]	Dementia	Blogs		TF-IDF, SUBTL	LR, NN
Hawkins[18]	Dreamers	Participants	LIWC		Manual
Niederhof.[3]	Dreams	DreamsCloud	LIWC	CDI	LDA
Oak[19]	Events	Twitter	SRILM	BLEU	n-grams
Reece[1]	Mental cond.	Twitter	MTurk, LIWC	BDI, CES-D	HMM
Goodman[8]	Development	Participants			RSA
Pool[21]	Emotions	ISEAR	Scikit-learn		BOW
Sérra[4]	Hate speech	Social networks	MTurk		RNN
Perez-R.[25]	Motivation	Participants	MTurk	ICC	Manual
Penneb.[2]	Acad. success	Participants	LIWC		Manual

6 Findings from included works

In the following, we will mainly focuses on firstly some important findings of the included work for the research questions, and secondly on granting a 'big picture' of a possible general connection between language and cognitive processes. An overview of the problem domains (without the approaches and data sources, as they are task specific), tools, psychometric measures and research methods can be found in Table 1.

6.1 Language and emotions

Hate speech detection has been a popular task ever since the recent discussion of verbal abuse on social networks has dominated some headlines [4]. Hate speech is especially prone to neologism, out-of-vocabulary words (OOV) and a lot of noise in the form of spelling and grammar mistakes. Furthermore, a known vocabulary of words that can be considered part of hate speech gets outdated rapidly. Serrà *et al.* [4] proposed a promising two-tier approach by training next character prediction models for each class as well as training a neural network classifier that takes said class models as input in order to measure the distance of expectation and reality. They achieved an accuracy of 0.951. Thus, in order to detect hate speech, it is more important to focus on *how* people alter their use of language rather than to focus on the particular words.

Dreams. Niederhoffer *et al.* [3] researched dream language by analyzing the content with an LDA topic model [15], categorizing emotions by the emotional classification model [10] and linguistic style via LIWC [16]. Dreams can be described as narratives, that predominantly describe past events in a first person point of view via first person pronouns with a particular attention to people, locations, sensations (e.g. hearing, seeing, the perceptional process of feeling). Since those dream narrations often exceed observations that are explainable by the dreamers (e.g. different physical laws of the observable world), complex cognitive processes can be assumed. Due to lexical categories revealing those connections, it can be concluded that it is more important *how* people express their dreams, rather than *what* they state.

Distress. In order to detect distress on Twitter, Homan *et al.* [5] asserted the so-called sadness score from LIWC together with keywords and could show a direct link to the distress and anxiety of Twitter users. Homan *et al.* [5] also analyzed the importance of expert annotators and showed that their classifier, trained with expert annotator labels, achieved an F-score of 0.64. This direct link adds to the impression, that the way people express themselves is connected to cognitive processes.

All of the above mentioned findings and their direct conclusions lead to an answer of the **research question i)** on the connection between emotions and language, which can be reacted upon with approval.

6.2 Cognitive performance and language

Works on subsequent academic success often induce strong biases such as the intuition that spelling mistakes indicate cognitive performance. The study of language and context that has been undertaken by Pennebaker *et al.* [2] indirectly tackles those biases, as the study targets a connection between the use of language and subsequent academic success by investigating college proposal essays with LIWC. Pennebaker *et al.* [2] could show that cognitive potential was not connected to *what* applicants expressed but rather to *how* they expressed themselves in terms of closed class words such as pronouns, articles, prepositions, conjunctions, auxiliary verbs or negation. However, correlations over four years of college measured each year, ranged from $r = 0.18$ to $r = 0.2$, which are significant, but not very high. The second **research question ii)** targets a connection between cognitive performance and the use of language. Closed class words such as function words have shown a connection with subsequent academic success. Therefore, the research question can be confirmed, that cognitive performance can be connected with the use of language.

6.3 Changing language and cognitive processes

As Goodman *et al.* [7] pointed out, many phenomena in natural language processing such as implication, vagueness, non-literal language are difficult to detect. Some aspects of the use of language even stay unnoticed by speakers themselves: at times the use of language on social media platforms indicate early staged physical or mental health conditions, which even holds true when the speaker is not yet aware of the health decline [1] him- or herself, which induces the importance for early detection via use of language. By using aspects of informed speakers and game theory, Goodman *et al.* [7] achieved a correlation of $r = 0.87$.

Reece *et al.* [1] were able to detect an early onset of dementia through tweets (Twitter posts) up to nine months before the official diagnosis of participants has been made (F1 = 0.651). Moreover, the word shift graphs of Hidden Markov Models (HMM) on time series in a sliding window could show a course of disease from early changes in language to stronger changes and a diagnosis until a normalization of language use as the condition was treated. This change has been detected by the labMT happiness score, which is a sentiment measurement tool for psychologically depicted scores on a dictionary, similar to LIWC. Thus, the connection of mental changes and the use of language, subject to **research question iii** can be confirmed as well.

7 Conclusion

Across most studies of included works, there are two main conclusion.

Reduction of bias. It has shown that function words can be the key factors of grasping the psyche of humans by surveying their use of language. Fine *et al.* [28] showed in their work that some corpora unknowingly induced strong

biases that alter the objectivity of said corpora – e.g. the corpus of Google for n-grams over-predicts how fast technological terms are understood by humans. Researchers tend to resort to strong biases when designing e.g. data collection for corpora or classification tasks, since experiences seemingly foretell e.g. cognitive performance with biased measures such as the usage of a complex grammar, eloquence or of making few spelling mistakes – as explained by Pennebaker *et al.* [2] –, thus leading us to the following, second conclusion:

Focus on *how* people express themselves, rather than *what* they express. The three research questions and their answers have lead to a hypothesis based on findings of included works: in order to grasp the psyche by the use of language, it is more important to survey *how* people express themselves rather than which words are actually used. Most important findings when looking at NLPsych have in common, that a possible key for accessing the psyche lies in small words such as function words or with dictionaries developed by psychologists that focus on cognitive associations with words rather than the lexical meaning of words, such as LIWC, labMT or ANEW. Furthermore, function words are more accessible, easier to measure and easier to count than e.g. complex grammar. Thus leading us to the conclusion that in order to access the psyche of humans through written texts, the most promising approaches are data driven, aware of possible biases and focus on function words rather than a content-based representation.

8 Future work

This section discusses some possible next steps for research in NLPsych.

A connection of scientific fields and sub-fields. As shown in Section 2, natural language processing in the sub-field of psychology is mainly about the study of language in clinical psychology and thus connected to mental conditions and diseases. Findings from other application areas such as dream language or the connection of language and academic success as indicators for cognitive performance could be valuable if connected to, or if used in other domains.

Researchers should rely more on best practice approaches. Some included work such as Reece *et al.* [1] demonstrate the advantages the sub-field can experience if state of the art methods are used and connected in order to access the full potential of natural language. As Morales *et al.* [9] pointed out, it is promising to enhance promising research approaches with state of the art and best practice methods, as well as a connection to other sub-fields for future development of a natural language processing.

Use established psychometrics combined with NLPsych. Whilst the already mentioned perceptions for future work – the connection of sub-fields and the usage of best practice approaches – are rather natural and known by many researchers, one possible research gap of NLPsych, the operant motive test (OMT) – developed by Scheffer *et al.* [37] –, illustrates the potential that NLPsych holds. The OMT is a well established psychometrical test that asserts the fundamental motives of humans by letting participants freely associate usu-

ally blurred images. Said images show scenarios in which labeled persons interact with each other. Participants are asked to answer questions on those images.

Since trained psychologists do not solemnly rely on provided word lists but rather develop an intuition for encoding the OMT – nonetheless showing high cross-observer agreement – that enables them to access the psyche, there has yet to be a method to be developed for this intuition by using best practice approaches and connecting scientific fields. This way, artificial intelligence might become even better at 'reading between the lines'.

References

[1] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer, "Forecasting the onset and course of mental illness with twitter data," *Nature Scientific Reports*, vol. 7, no. 1, p. 13 006, 2017, ISSN: 2045-2322. DOI: 10.1038/s41598-017-12961-9. [Online]. Available: https://www.nature.com/articles/s41598-017-12961-9.

[2] J. W. Pennebaker, C. K. Chung, J. Frazee, G. M. Lavergne, and D. I. Beaver, "When small words foretell academic success: The case of college admissions essays," *PLOS ONE*, vol. 9, no. 12, e115844, 2014, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0115844. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115844.

[3] K. Niederhoffer, J. Schler, P. Crutchley, K. Loveys, and G. Coppersmith, "In your wildest dreams: The language and psychological features of dreams," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, Vancouver, BC, Canada: Association for Computational Linguistics, 2017, pp. 13–25. [Online]. Available: http://www.aclweb.org/anthology/W17-3102.

[4] J. Serrà, I. Leontiadis, D. Spathis, G. Stringhini, J. Blackburn, and A. Vakali, "Class-based prediction errors to detect hate speech with out-of-vocabulary words," in *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 36–40. [Online]. Available: http://www.aclweb.org/anthology/W17-3005.

[5] C. Homan, R. Johar, T. Liu, M. Lytle, V. Silenzio, and C. O. Alm, "Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Baltimore, MD, USA: Association for Computational Linguistics, 2014, pp. 107–117. [Online]. Available: http://www.aclweb.org/anthology/W14-3213.

[6] M. Tomasello, *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, 2nd. NJ, USA: Psychology Press, 2002, 376 pp., ISBN: 978-1-317-69352-9.

[7] N. D. Goodman and M. C. Frank, "Pragmatic language interpretation as probabilistic inference," *Trends in Cognitive Sciences*, vol. 20, no. 11, pp. 818 –829, 2016.

[8] J. C. Goodman, P. S. Dale, and P. Li, "Does frequency count? Parental input and the acquisition of vocabulary," *Journal of Child Language*, vol. 35, no. 3, pp. 515–531, 2008, ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000907008641. (visited on 03/21/2018).

[9] M. Morales, S. Scherer, and R. Levitan, "A cross-modal review of indicators for depression detection systems," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, Vancouver, BC, Canada: Association for Computational Linguistics, 2017, pp. 1–12. [Online]. Available: http://www.aclweb.org/anthology/W17-3101.

[10] G. Coppersmith, K. Ngo, R. Leary, and A. Wood, "Exploratory analysis of social media prior to a suicide attempt," presented at the CLPsych 2016, San Diego, CA, USA, 2016, pp. 106–117. DOI: 10.18653/v1/W16-0311. [Online]. Available: http://www.aclweb.org/anthology/W16-0311.

[11] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses - semantic scholar," presented at the CLPsych@HLT-NAACL, Denver, CO, USA, 2015. [Online]. Available: www.aclweb.org/anthology/W15-1201.

[12] R. Kshirsagar, R. Morris, and S. Bowman, "Detecting and explaining crisis," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, Vancouver, BC, Canada: Association for Computational Linguistics, 2017, pp. 66–73. [Online]. Available: http://www.aclweb.org/anthology/W17-3108.

[13] J. H. Shen and F. Rudzicz, "Detecting anxiety on Reddit," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, Vancouver, BC, Canada: Association for Computational Linguistics, 2017, pp. 58–65. [Online]. Available: http://www.aclweb.org/anthology/W17-3107.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, ISSN: 0885-6125. DOI: 10.1023/A:1022627411411. [Online]. Available: https://doi.org/10.1023/A:1022627411411.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003, ISSN: 1532-4435. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937.

[16] J. Pennebaker, C. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The development and psychometric properties of LIWC2007," Software Manual, Austin, TX, USA, 2007.

[17] V. Masrani, G. Murray, T. Field, and G. Carenini, "Detecting Dementia through retrospective analysis of routine blog posts by bloggers with Dementia," in *BIONLP 2017*, Vancouver, BC, Canada: Association for

Computational Linguistics, 2017, pp. 232–237. [Online]. Available: http://www.aclweb.org/anthology/W17-2329.

[18] R. Hawkins and R. Boyd, "Such stuff as dreams are made on: Dream language, LIWC norms, & personality correlates," *Dreaming*, vol. 27, 2017. DOI: 10.1037/drm0000049.

[19] M. Oak, A. Behera, T. Thomas, C. O. Alm, E. Prud'hommeaux, C. Homan, and R. Ptucha, "Generating clinically relevant texts: A case study on life-changing events," *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pp. 85–94, 2016. DOI: 10.18653/v1/W16-0309. [Online]. Available: https://aclanthology.info/papers/W16-0309/w16-0309.

[20] I. Lancashire and G. Hirst, "Vocabulary changes in Agatha Christie's mysteries as an indication of Dementia: A case study," in *Cognitive Aging: Research and Practice*, ser. Cognitive Aging: Research and Practice, Toronto, Canada, 2009, pp. 8–10.

[21] C. Pool and M. Nissim, "Distant supervision for emotion detection using Facebook reactions," in *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, Osaka, Japan, 2016, pp. 30–39. [Online]. Available: https://aclanthology.info/papers/W16-4304/w16-4304.

[22] D. Benikova, M. Wojatzki, and T. Zesch, "What does this imply? Examining the impact of implicitness on the perception of hate speech," in *Language Technologies for the Challenges of the Digital Age*, G. Rehm and T. Declerck, Eds., Cham: Springer International Publishing, 2018, pp. 171–179, ISBN: 978-3-319-73706-5.

[23] W. Warner and J. Hirschberg, "Detecting hate speech on the World Wide Web," in *Proceedings of the Second Workshop on Language in Social Media*, Montreal, Canada: Association for Computational Linguistics, 2012, pp. 19–26. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390374.2390377.

[24] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1–10. DOI: 10.18653/v1/W17-1101. [Online]. Available: http://aclweb.org/anthology/W17-1101.

[25] V. Pérez-Rosas, R. Mihalcea, K. Resnicow, S. Singh, and L. An, "Building a motivational interviewing dataset," *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pp. 42–51, 2016. DOI: 10.18653/v1/W16-0305. [Online]. Available: https://aclanthology.info/papers/W16-0305/w16-0305.

[26] M. Wolf, A. Horn, M. Mehl, S. Haug, J. Pennebaker, and H. Kordy, "Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count," *Diagnostica*, vol. 54, pp. 85–98, 2008. DOI: 10.1026/0012-1924.54.2.85.

[27] R Baayen, R Piepenbrock, and H. Rijn, *The CELEX lexical data base [CD-ROM Manual]*. Pennsylvania, USA: Linguistic Data Consortium, University of Pennsylvania, 1993.

[28] A. Fine, A. F. Frank, T. F. Jaeger, and B. Van Durme, "Biases in predicting the human language model," in *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, vol. 2, Baltimore, MD, USA, 2014, pp. 7–12. DOI: 10.3115/v1/P14-2002.

[29] A. Stolcke, "SRILM — an extensible language modeling toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, vol. 2, Denver, CO, USA, 2004.

[30] 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 Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, ISSN: 1532-4435. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2078195 (visited on 04/24/2018).

[31] R. N. Jørgensen, P. S. Dale, D. Bleses, and L. Fenson, "Clex: Across-linguistic lexical norms database*," *Journal of Child Language*, vol. 37, no. 2, pp. 419–428, 2010, ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000909009544.

[32] M. Võ, M. Conrad, L. Kuchinke, K. Urton, M. Hofmann, and A. Jacobs, "The Berlin affective word list reloaded (BAWL-r)," *Behavior research methods*, vol. 41, pp. 534–8, 2009. DOI: 10.3758/BRM.41.2.534.

[33] *Psychometric*, in *The Free Dictionary*, M. O Toole, Ed., 2018. [Online]. Available: https://medical-dictionary.thefreedictionary.com/psychometric.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of Workshop at ICLR*, Scottsdale, AZ, USA, 2013.

[35] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *ARXIV:1406.1078 [cs, stat]*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734. arXiv: 1406.1078. [Online]. Available: http://www.aclweb.org/anthology/D14-1179.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[37] D. Scheffer and J. Kuhl, "Der Operante Motiv-Test (OMT): Ein neuer Ansatz zur Messung impliziter Motive," in *Tests und Trends, Jahrbuch der psychologischen Diagnostik*, J. Stiensmeier and F. Rheinberg, Eds., vol. N.F.2, Göttingen, Germany: Hogrefe Verlag, 2003, pp. 129–150.