Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1178–1190
Marseille, 11–16 May 2020

Stigma Annotation Scheme and Stigmatized Language Detection in Health-Care Discussions on Social Media

Nadiya Straton, Hyeju Jang, Raymond Ng
Center for Business Data Analytics, Data Science Institute, Data Science Institute
Copenhagen Business School, University of British Columbia, University of British Columbia
ns.digi@cbs.dk, hyejuj@cs.ubc.ca, rng@cs.ubc.ca

Abstract

Much research has been done within the social sciences on the interpretation and influence of stigma on human behaviour and health, which result in out-of-group exclusion, distancing, cognitive separation, status loss, discrimination, in-group pressure, and often lead to disengagement, non-adherence to treatment plan, and prescriptions by the doctor. However, little work has been conducted on computational identification of stigma in general and in social media discourse in particular. In this paper, we develop the annotation scheme and improve the annotation process for stigma identification, which can be applied to other health-care domains. The data from pro-vaccination and anti-vaccination discussion groups are annotated by trained annotators who have professional background in social science and health-care studies, therefore the group can be considered experts on the subject in comparison to non-expert crowd. Amazon MTurk annotators is another group of annotator with no knowledge on their education background, they are initially treated as non-expert crowd on the subject matter of stigma. We analyze the annotations with visualisation techniques, features from LIWC (Linguistic Inquiry and Word Count) list and make prediction based on bi-grams with traditional and deep learning models. Data augmentation method and application of CNN show high performance accuracy in comparison to other models. Success of the rigorous annotation process on identifying stigma is reconfirmed by achieving high prediction rate with CNN.

Keywords: Annotation process, Stigma Annotation Scheme, Social media, CNN, N-grams.

1. Introduction

In this paper, we study *stigma* – we aim to distinguish stigmatised language from non-stigmatised using machine learning and natural language processing (NLP). We formulate the problem as text classification, and explore different classification algorithms in order to understand the components of "health stigma" better.

Stigma is a concept related to bias, prejudice, and stereotype. According to the definition in Merriam-Webster (2011), *stigma* is a mark of shame or discredit, while *stereotype* is a standardized mental picture that is held in common by members of a group and that represents an oversimplified opinion, prejudiced attitude, or uncritical judgment. *Bias* a personal and unreasoned judgment, and *prejudice* is preconceived opinion, formed without just grounds or before sufficient knowledge, an irrational attitude of hostility directed against an individual, a group, a race, or their supposed characteristics. Although these concepts are slightly different, they are all closely related in that all these negative attitudes can lead to discrimination. In the current study, we will use stigma, bias, prejudice, and stereotype interchangeably.

Stigma is a barrier to quality healthcare, which often undermines health seeking behaviors like diagnosis and engagement in treatment. Katz (2014) and Joseph et al. (2015) observe that there may be an unconscious tendency to assign blame and stigmatize against health-care conditions that seem more threatening and unknown than against health-care conditions that are as dangerous but better understood. Furthermore, Stuber et al. (2008), Saguy (2012), and De Brún et al. (2014) imply that those attitudes can hinder the stigmatised person/group from seeking medical attention and professional advice outside of their close circle and thus contributes to health disparities.

The importance of stigma has drawn attention to identification of stigma in social media texts (Reavley and Pilkington, 2014; Li et al., 2018), where the authors follow a framework tailored to identify depression stigma. However, to the best of our knowledge, there is no rigorous study on stigma annotation or annotated corpus available, which hinders development of a large scale study that leverages contemporary machine learning.

In this paper, we study stigma about immunization on social media. We investigate discussions around pro-vaccination and anti-vaccination, and develop the annotation scheme for stigma based on social science theories. We then perform a corpus study on the data from Facebook groups on vaccination. In addition, we explore various deep learning models and traditional machine learning models to identify stigma in the text. Analyses of features from annotated corpus and conclusions made in this paper will be beneficial for future studies on stigma identification not only in the vaccination context, but across diverse health-care conditions.

The following research questions will be addressed in the paper:

Q1: How to build a rigorous annotation scheme and achieve higher inter-rater agreement when there is no consensus on a concept definition among the researchers?

Q2: What are the characteristic features of stigmatised language in vaccination comments on social media?

Q3: Can deep learning models be better predictors of health stigma given the relatively small labelled dataset?

2. Literature Review

Three main bodies are introduced in this section: computational models in pro-vaccination and anti-vaccination discourse on social media, computational models of stigma in

health and social media, and stigma annotation.

2.1. Quantitative studies on pro-vaccination and anti-vaccination discourse on social media

Computational modeling of behaviour in relation to immunization on social media is a relatively recent research topic, but has started to receive more attention in the last few years. For this review, we searched for keywords, "vaccination" and "social media" in *Science Direct*, *Scopus*, and *PubMed* databases. Only sources from 2006 to the beginning of 2019 were taken into account. Initial screening of the title for relevance resulted in 57 research papers. After abstract review, removal of duplicates and irrelevant records, 25 articles were selected for manual review and obtained in full text. Furthermore, the selection was narrowed down to nine quantitative studies, where from the data were abstracted on study characteristics, research aim, identification of relevant features, and the type of evaluation performed. Most computation work on vaccinations on social media can be classified into studies that investigate polarization from the information perspective, together with several additional features (Schmidt et al., 2018; Mitra et al., 2016; Pennebaker et al., 2001; Tomeny et al., 2017; Faasse et al., 2016; Du et al., 2017; Massey et al., 2016; Kang et al., 2017), and studies that focus on behavioural patterns from the user/community networks perspective with their inherent temporal trends (Bello-Orgaz et al., 2017; Kostkova et al., 2017). Discussed research is based on Facebook or Twitter data exclusively, with no cross-platform application. Schmidt et al. (2018) concluded that Facebook contributes to polarized attitudes as a powerful promoter of different sentiments about vaccinations and therefore contributes to vaccine hesitancy. Mitra et al. (2016) describe polarisation in attitudes of those who persistently hold anti-vaccination sentiment over a long term, show government distrust, and general paranoia and those who hold pro-vaccination attitudes, with opposite sentiment. However new adopters of polarized sentiments can be easier swayed in either direction. Tomeny et al. (2017) suggested that polarization stems from demographic characteristics, which partially explains in-group formation of polarized opinions. Kostkova et al. (2017) suggested influential groups / people to influence the health-care behaviour and opinion through social network. Latter confirms that the decision to vaccinate or not depends on public opinion, which can be shaped by a successful communication strategy among the in-group members or new adopters of anti-vaccination attitudes who are easier to convince.

2.2. Computational models of stigma in health and social media

One of two studies on identification of stigma in social media aimed at detecting depression stigma using linguistic features (Li et al., 2018). Li et al. (2018) established two tasks: differentiation between Chinese Weibo posts with or without depression, and differentiation among three specific types of depression stigma. Study might be difficult to generalize due to imbalanced data, as only *6%* of the posts indicated depression stigma and only one social media source was used (Li et al., 2018). Reavley and Pilkington (2014) coded Twitter posts into different sub-themes, including mental illness category. Their findings resulted in *0.7%* tweets (43 tweets) annotated as stigma, whereas the rest had neutral or positive sentiment. Coding frameworks in both studies were primarily tailored to depression stigma and therefore cannot be generalized to other health-care conditions or domains. Current work aims to add a new layer to the foundation on computation models of stigma, and introduce an annotation scheme that can be applied across different health-care domains.

2.3. Stigma annotation

Annotating stigma/prejudice/stereotype is a nontrivial task as there is a lack of annotation schemes and consensus from the researchers on the definition of stigma (Link and Phelan, 2001). Initial challenge in the current research was due to the lack of annotated datasets on the stigma topic. Therefore, the main studies of reference were annotation of metaphor and non-literal language that served as a foundation for designing the current annotation scheme (Jang et al., 2014; Group, 2007; Shutova and Teufel, 2010; Wallington et al., 2003). Wallington et al. (2003) in their study on metaphor annotation instructed one group of data collection assistants to annotate an "interesting stretch" of text, as "metaphor" that was too abstract to interpret. Stigma, similarly to metaphor, is notoriously difficult to identify.
Prior work of Link and Phelan (2001) draw attention to the *"definition and utility of the stigma concept and that none of the conceptualizations should be viewed as definitive"*. Authors elaborate that the concept has been applied to a wide array of circumstances that are likely to lead to differences in the interpretation (Link and Phelan, 2001). Research in the area is multidisciplinary in its nature, where several fields within social science contribute with their own approaches and views on what is inherent to the concept (Link and Phelan, 2001). To improve on existing annotation schemes, the complex definition based on theoretical frameworks from (Goffman, 2009; Allport et al., 1954; Link and Phelan, 2001) was split into simpler definitions centered around conceptual characteristics similarly to the study by (Wallington et al., 2003), where the process was systematised into a number of steps and split into simpler concepts. Splitting the stigma definition into simpler concepts was adopted in the current research as well. However, the above mentioned approaches do not distinguish between different degrees of the concept, whereas the present work introduced several layers of the concept that were reduced to two in the process: blame and out-of-group generalization.

3. Data

In this study, a novel health-care dataset from the two biggest health-care walls on Facebook that discuss vaccines is introduced for the detection of stigma. One is focused on pro-vaccination (*Refutations to Anti-Vaccine Memes*, with about 283,274 followers) and the other on anti-vaccination (Dr. Tenpenny on *Vaccines and Current Events*, with about 224,851 followers). Most subscribers to each wall have homogeneous in-group opinions on the topic, with limited or no reservation to express their opinions freely and are

Table 1: Data Characteristics.

Datasets	# Comments	# Sentences	# Words	# Tokens	# Types
Anti-Vaccinations	1,468	2,978	31,348	35,780	4,805
Pro-Vaccinations	1,293	2,173	22,359	25,515	3,659
Total	2,761	5,151	53,707	61,295	8464
Anti-Vaccination Balanced	2,556	4,376	38,689	44,564	4,805
Pro-Vaccination Balanced	2,453	3,718	34,652	39,788	3,659
Total	5,009	8,094	73,341	84,352	8,464

likely to receive support from the group members. This type of discussion within a community serves as an ideal dataset for stigma/prejudice/stereotype identification, because it is highly polarized with connotations of intolerance to the opinions of the out-of-group members carrying different point of view. A total of 4,502 comments (2,251 anti- and 2,251 pro-vaccination comments, 8,584 sentences, and 105,470 tokens) were collected from January to March 2018.

To annotate the data, each comment was labelled by trained annotators and Amazon MTurk experts three times according to class definitions: *stigma*, *not stigma*, and *undefined*. Non-consensus comments were removed in two steps with final "Golden Standard" dataset containing 2,761 comments, 5,151 sentences, and 61,295 tokens, as shown in Table 1. Detailed description of the process on the removal of non-consensus posts and the purpose is described in Section 4. Study Design, Subsection 4.1.3. There were more stigmatized comments than not stigmatized and undefined comments (In pro-vaccination discussions 63,34% of comments contain stigma, 25,06% do not contain stigma sentiment, and 11,06% were defined as neither. In anti-vaccination discussions - 58,04% of comments contain stigma, 30,79% do not contain stigma sentiment, and 11,017% labelled as undefined). Therefore, the latter two classes were padded with additional data points from each of the respective labelled categories - up-sampled to achieve a more balanced dataset with even share of labels, each label equaling roughly one third of the data points.

4. Study Design

.

The main goals of our current work are: 1) design the annotation scheme that can be applied in other health-care domains, 2) improve annotation process that results in higher inter-rater agreement, and 3) extract a feature set and build computational models for stigma classification.

In the process of annotation we define three annotation categories: *stigma*, *not stigma*, and *undefined*. The following examples show characteristic messages within each of the categories for anti-vaccination messages.

1. Stigma
    - *"IOW pharma is a bunch of criminals who will have him whacked."*
    - *"docs cannot survive on 15 medicais payments per patient, so they are forced to pimp their soul for dollars"*
2. Non-stigma
    - *"YouTube deleted the health rangers entire page and over 1700 videos"*
3. Unknown/Undefined
    - *"Must be that new strain of it that's floating around."*

In the anti-vaccination comments labelled as stigmatized, blame is placed on an out-of-group, which is primarily a government organization or an institution. *"IOW pharma is a bunch of criminals"* shows conflict, fear, strong negative emotion, and animosity towards an institution. Opposing an out-of-group and hostility towards it helps to maintain and keep in-group membership and conformity more stable (Allport et al., 1954).

Stigmatised language is also expressed through inflexible and unfounded generalization, projection, and unsupported judgement in the following comment: *"docs cannot survive on 15 medicais, so they are forced to pimp their soul for dollars"*. The statement shows a widely held but a fixed and oversimplified image or idea of a particular type of person or thing (Link and Phelan, 2001). If people are judged as out-of-group members, the perceiver will see them as especially similar and lacking in variability (Fiske, 1998). On the contrary, the non-stigmatized statement, *"YouTube deleted the health rangers entire page"* does not contain prejudice. It is most likely based on a fact.

Undefined/unknown statements usually require context to make a decision: *"Must be that new strain of it that's floating around"*. It is unclear what is meant by *"new strain of it"* without further elaboration.

Examples of pro-vaccination messages that belong to the three categories are below.

1. Stigma
    - *"All because of a few idiotic parents that should be in jail."*
    - *"Anti-vaccine parents control their kids...until they grow up."*
2. Not Stigma: *"I love science!"*

3. Unknown/Undefined: *"I cant even"*

In pro-vaccination comments, antagonism is clearly directed at their out-of-group: the anti-vaccination movement and parents. Blame and conflict are observed in *"all because of a few"*, which at the same time points to generalization and condescending feelings, whereas *"should be in jail"* shows personal unsupported judgement. Stigma occurs when elements of labeling, stereotyping, cognitive separation into categories of *"us"* and *"them"* happen in the situation of power or authority that allows these components to unfold easier (Link and Phelan, 2001). In addition, Fiske (Fiske, 1998) points out that comparable to the group-level threats, specific out-of-group members are presumed to block in-group goals and the rhetoric is supported with over-generalised statements, and unsupported judgement, as can be observed in the following statements: *"Anti-vaccine parents control their kids"*.

In contrast, *"I love science!"* expresses one's state at the time of writing the comment, and is most likely a fact. Therefore, it is assigned to the *"Not Stigma"* category.

"I cant even" connotation can be interpreted in different ways, as it lacks the context to make a decision. Therefore, it assigned to the unknown category.

4.1. Annotation process

The annotation process on the anti-vaccination dataset started with a pilot test annotating 100 randomly selected anti-vaccination posts. Six annotators labelled each comment before they read the initial annotation instruction on *stigma* and *not stigma* categories (each comment was annotated six times). We measured inter-agreement among annotators by computing Cohen's kappa (Cohen, 1960) of each pair of annotators. The distribution of pair-wise κ fell in the range of $[12.4-36.6]$ prior to annotators reading the annotation guidelines. After reading the initial annotation guidelines and training on category definition the Cohen's Kappa did not show significant improvement. It was clear that annotation instruction needed to be improved to include more precise definitions and less annotation labels (see Table 7 in Appendix A). With new instructions the range of Cohen's Kappa κ slightly improved to $[16-50]$. The initial result showed that, despite the small sample size, annotators who are linked to the same country and those who have similar political convictions achieved a higher inter-rater reliability rate. During the labeling process, annotators underlined words and expressions which were deemed to be stigmatising, and provided their interpretation. Hence, the annotation scheme was gradually updated with new definitions before the entire corpus was annotated.

4.1.1. Annotation Scheme

Pilot test was based on the research literature that studied stigma, prejudice, bias, and stereotype concepts (Allport et al., 1954), (Goffman, 2009), (Fiske, 1998), (Link and Phelan, 2001). Initial annotation scheme included twelve labels that were grouped based on the concepts that appeared frequently in the literature. Those labels were: 1) Placing blame and lacking evidence, 2) Sign of conflict, 3) Expressed hostility and aggression, 4) Ascription of materialistic, impersonal or inhumane values, 5) Strong sentiment, emotion, and lack of evidence, 6) Irrational judgement, 7) Expressed avoidance without reason, 8) Exaggeration, fear and paranoia, 9) Not founded suspicion of hidden agenda, 10) Generalising and assigning one trait to a person, group, or everyone - "He/she/they" followed by "always, all the time, all", 11) Homogenising a problem based on a single or a small number of instances, and 12) Predicting, guessing, presuming, or projecting hypothetical scenarios. In the process of labelling annotators suggested Personal opinion category and Opinions influenced by mass media or celebrity. Multitude of labels was not a purpose in itself but rather aimed at exhaustive description of the concept. However, it was also observed that some categories overlap for example: 1) Placing blame, 6) Irrational judgement, and 9) Suspicion of hidden agenda; 5) Strong emotion/sentiment and 8) Exaggeration. Moreover, too many categories confused annotators, rather than supported in identifying the sentiment. Annotation scheme was modified to include nine categories with more precise and concise description and comment/post examples (see Table 7 in Appendix A). Inter-rater agreement was still rather low, but served as an indication that improved annotation instruction lead to higher inter-rate agreement. Based on the underlined comments, further analyses and discussion with annotators resulted in an annotation scheme that contained four categories and two levels of stigma: 1) Expressions that sustain hostility - sentiment focused on blame or antagonism, 2) Expressions that sustain inconsistency and over-generalization - out-of group attitudes, 3) Lacking context to make a decision, and 4) Not stigma (see Table 2). Allport et al. (1954) refers to "scapegoat", "whipping-boy" while Goffman (2009) mentions tainted, discounted person. Fiske (1998) also suggests that people confuse other people by lumping them into the same discounted category.

Prejudgements and misconceptions become prejudices only if they are not reversible when exposed to new knowledge, however they were treated as synonyms to stereotype, prejudice, bias, and stigma. When stereotyped sentiment is mixed, such as both hostility and over-generalization are present, annotators were asked to use a judgement to choose the stronger matching sentiment. Lacking context to make decision category meant potentially ambiguous content. It was not clear from the text if comment contains stigma, prejudice, stereotype or is void of any of the mentioned sentiments. Not stigma category - does not contain stigma/prejudice/bias/stereotype, is based on a fact or personal experience, is void of inflexible generalization and stigmatised sentiments prevalent in categories 1 and 2.

4.1.2. Annotation by Trained Annotators

After annotation instructions were updated (Table 2), a total of 2,251 anti-vaccination comments were annotated by trained annotators, each comment was annotated three times by a different annotator. All thirteen annotators were recruited through personal networks, and had never performed an annotation task before. All annotators were fluent in English, represented diverse age groups and were primarily from social science or health-care educational back-

Table 2: Annotation scheme - 4 labels

Does the sentiment convey stereotype/prejudice/bias/stigma? If YES option 1-2, if lacking context - 3, if NO then 4 (NONE)	
Label	**Post/Comment Example**
1. Expressions that sustain hostility: Blame, Suspicion, Conflict (Hate, Fear), Exaggeration, Strong emotion, An insult, Rejection, Animosity, Condescension, Aggression.	".. NASA faked the moon landings do you really want to believe anything they say? NASA steals 50 million USD a DAY and all we get is really bad CGI and "space bubbles". "They may be injecting mothers before babies are born so they can say the baby was born with autism and it is genetic." "..No kidding, it's cuzz they don't protect you from anything, govt way to inject poison called population control!" "Boycott his movies", "He's a narcissistic coke head, "This guy is a stick SOBb", "Antichrist", "Evil pure evil !"
2. Expressions that sustain Inconsistency and Overgeneralization: Inflexible Unfounded Over-Generalization, One-sided interpretation, Predicting, Guessing, Unsupported judgement, Personal opinion, Dichotomization, Tabloid thinking, Demagoguery.	".. only Korea and Japan have good ones they are smart and know how to manage them but they have no illegals or refugees either those ruin healthcare they do not pay into it." "... we don't have any rights, they bully us into it by slapping a mask on, when there's no medical or scientific evidence of the mask preventing it if you're asymptomatic, but the hospitals have to meet 90% compliance in order to get reimbursed by medicare thanks to our great health care system" "..I can guarantee the brainwashed will be flocking to get the jab! They've probably not made as much money on the flu jabs this year.. Scaremongering!" "They are definitely spraying something." "There are only two kinds of people: the weak and the strong"
3. Lacking context to make a decision	"So, it's a success." "??" "I'm a little confused. I thought Kennedy wasn't for forced vaccinations." "My son's private school which I'm not going to mention there is a person who is undercover who is trained and carries a gun.... Maybe one day this will just be the norm metal detectors undercover officers."
4. Not stigma	"Been saying this for years" "This is the worst thing I have ever heard of." "How do they know the bug is Influenza virus?"

ground. More detailed annotator profiles are available in Table 3. Two annotators were reimbursed for the annotation task while eleven volunteered. Annotators were guided on the meaning of each category in the context of vaccination comments on Facebook before and during the annotation process, however performed annotations independently of each other.

4.1.3. Annotation by Amazon MTurk annotators

To compare the agreement rate between trained annotators, we also annotated anti-vaccination comments with MTurk experts. All MTurk annotators were required to be "masters" to do the task, demonstrated excellence through a wide range of tasks, with HIT approval rate (%) greater than *99%*. No further premium qualifications such as education level, social science background, or years of experience were required or known at the time. We allotted one minute per comment, and provided $0.05 reward per assignment.

Each comment, otherwise referred to as 'markable', was annotated three times by a set of coders (annotators) (c), who assigned labels from a set of categories (k) - labels (Artstein and Poesio, 2008). We then computed observed agreement Ao (Artstein and Poesio, 2008), which measures the percentage of judgements on which the annotators agree when coding the same data independently (divided by the total number of items). See below for the equation.

$$Ao = \frac{1}{i}\sum_{i \in I} arg_i, \quad \arg_i \text{ for all items } i \in I$$

where:

$$\arg_i = \begin{cases} 1 & \text{if the three coders assign i to the same category} \\ 0 & \text{if the three coders assign i to different categories} \end{cases}$$

A higher percentage of agreement was achieved through elimination of comments in two rounds:

1. Non consensus removed 1:

Table 3: Annotator Profiles.

Background (years in labour force)	Education	Age	Gender	English as a first language	Country of residence	Volunteer
History student	High school, courses at UBC	21	m	Yes	Canada	No
Social science	Bachelor's degree	38	f	No	UK	No
X-ray technician	Diploma	Retired	f	Yes	Canada	Yes
Teacher (retired)	Master degree in education	N/A	f	Yes	Canada	Yes
Nurse	Bachelor's degree in nursing science	N/A	f	Yes	Canada	Yes
Medical clerical assistant (20)	High school	N/A	m	Yes	Canada	Yes
X-ray technician (12)	Diploma, technology in medical radiation	N/A	f	Yes	Canada	Yes
X-ray technician (25)	Diploma, technology in medical radiation	N/A	m	Yes	Canada	Yes
Psychology	Bachelor's degree	N/A	f	Yes	Canada	Yes
Counselor, resident care (10)	Diploma	N/A	m	Yes	Canada	Yes
Sales and marketing (20)	Bachelor's degree sociology / psychology	N/A	m	Yes	Canada	Yes
Legal	Bachelor's in law	22	f	Yes	UK	Yes
Legal (15)	Master's degree in law	37	m	No	Denmark	Yes

- Three annotators disagreed on the category and assigned three different labels

2. Non consensus removed 2:

   - Opposite categories are assigned: one/two annotators assign stigma/not stigma and the rest assign opposite category
   - Three annotators agree and assign unknown/undefined category

The annotation practice measuring reliability with only two coders is seldom considered enough (Artstein and Poesio, 2008), as annotators might introduce their bias, thus eleven annotators were recruited, and roughly the same number of MTurk experts performed the annotation task independently, assigning three labels for each comment. Fleiss Kappa (Artstein and Poesio, 2008) was applied as done in (Fleiss, 1971) to measure "chance agreement" that reflects the combined judgments of all coders:

$$P(k) = \frac{1}{ic} n_k$$

where $P(k)$ is the expected agreement, i is the total number of assignments, c is the number of coders, n_k is the number of times an item i is classified in category k.

Results from trained annotators did not show a higher agreement rate than MTurk experts (see Table 4), therefore pro-vaccination dataset was annotated by MTurk experts only. Further combining two levels of stigma: 1) Expressions that sustain hostility and 2) Expressions that sustain inconsistency and over-generalization into one stigma category, resulted in the dataset with three categories: *stigma*, *not stigma*, and *undefined* and higher agreement rate. Two categories were formed by re-assigning *undefined* category into either *stigma*, *not stigma* label. If two annotators assigned *stigma* category and third assigned *undefined*, then *undefined* label was re-assigned to *stigma* category. If two annotators assigned *not stigma* category and third assigned *undefined*, then similar logic applied. However, when two annotators assigned *undefined*, then one would be re-assigned to *stigma* category and another to *not stigma* and majority voting would be decisive in assigning final label.

"Golden standard" dataset with 3 categories: *stigma*, *not stigma*, *undefined* and highest Fleiss Kappa of *0.62* - anti-vaccination and *0.54* - pro-vaccination context after non-consensus comments were removed (see Table 4) is used for further analyses and feature extraction. Share of agreement rate between three annotators consequently is *68%* for anti-vaccination comments and *62%*) for pro-vaccinaton comments.

4.2. Data Visualization

Data visualization analyses is based on "Golden standard" dataset before up-sampling. To see the relationships between class labels and features as in Figure 1 in Appendix B, correlations between each pair of features is measured with Chi-square, using the top 25 most frequent words, with $TF = 5$ per document frequency $DF = 1$. Size of the "bubble"in Figures in Appendix B represents the frequency of each word. Anti-vaccination rhetoric within the stigma category is centered around "*Big pharma*", "*government*", "*control*", and "*evil*", whereas the not stigma category contain more "*love*", "*thank*", "*family*", "*watch*", and "*news*"words. The unknown category is positioned between "stigma" and "not stigma" with words that can point towards stigmatized context such as "*shot*", "*medicine*", and "*poison*" and neutral or positive words such as "*good*", "*sad*", and "*hear*". Similarly, pro-vaccination rhetoric within stigma category is based on negative sentiment, blame and death references, such as "*idiot*", "*stupid*", "*want*", "*people*", "*die*" and Not stigma category pointing towards more positive or neutral words such as "*year*", "*school*", "*month*", "*daughter*", "*free*". Figures were generated using KH Coder (Higuchi, 2016), a toolkit for text visualization.

Table 4: Fleiss Kappa.

Share of comments with mutual agreement between 3 annotators	4 labels	3 labels	2 labels	# of comments
Trained Annotators Anti-Vaccination	0.18	0.25	0.33	2 251
Trained Annotators Anti-Vaccination Non Consensus Removed1	0.23	0.32	0.44	2 043
Trained Annotators Anti-Vaccination Non Consensus Removed2	0.37	0.53	0.81	1 351
MTurk Anti-Vaccination	0.25	0.34	0.57	2 251
MTurk Anti-Vaccination Non Consensus Removed1	0.33	0.44	0.66	1 992
MTurk Anti-Vaccination Non Consensus Removed2	0.42	0.62	0.84	1 468
MTurk Pro-Vaccination Dataset	0.18	0.24	0.36	2 251
MTurk Pro-Vaccination Non Consensus Removed1	0.23	0.32	0.49	1 955
MTurk Pro-Vaccination Non Consensus Removed2	0.35	0.54	0.82	1 293

4.3. Feature Extraction

Feature extraction analyses is conducted on the "Golden standard" dataset, with three annotation categories. Linguistic, grammatical, and psychological features were extracted according to the list from LIWC (Pennebaker et al., 2015). Z-score (standard error of the mean) computed for each selected feature (Table 7 in Appendix C) showed strength of sentiment for different categories. To understand if the reaction in a comment has stigma connotation, negative and positive emotion features might not be sufficient for the task. Allport in (Allport et al., 1954) referred to exploitation theory, where exploiting class for the purpose of stigmatizing shows one group as inferior and (Goffman, 2009) describes the inferior class as a second class citizens that bear the mark. Furthermore, *"justifying our own state of mind by reference to imagined intentions and behaviour of others"* has its roots in anxiety and underlying insecurity according to (Allport et al., 1954). Therefore additional features (i.e. clout, power, emotional tone, and anxiety) will help to computationally validate those theories against the annotated corpus.

4.4. Prediction

In this work, we compared various models presented in Table 5.

For all of the mentioned neural networks, the embedding size of the first layer was set to 100. We determined the hyperparameters empirically: RMSprop as the optimiser; learning rate was set to 0.001; drop out rate was set to 0.5. Training of the LSTM, BiLSTM, and CNN was done with 10 epochs and/or stopped early if the validation loss did not decrease after a fixed number of iterations. Dataset was randomly split on 80% training and 20% testing set. Embedding dimensions in BiLSTM were set to 100, maximum lenth to 100, rnn units to 1024, batch-size to 10.

5. Results

Annotation of social media comments for a person who rarely engages in social media can be quite tricky due to out-of-context phrases, abbreviations, and net-speak content. Representative annotation cohort was achieved by including volunteers who never used social media and those who use it frequently and of different age groups from *21* to *78*. Lower agreement rate between trained annotators in comparison to MTurk annotators might be surprising, as good portion of the "training" time is dedicated to explaining the annotation scheme and its definitions, purpose of the project, and how annotation results will be used in the research. Outsourced annotation tasks on MTurk entail concise explanation of the definitions and short annotation time. Higher agreement rate can be explained by experience in conducting annotation tasks by MTurk experts in contrast to trained annotators who never annotated a dataset before.

There are visible differences in the interpretation of the two types of stigma categories; the generalization sentiment proved to be the most difficult to detect. The boundary between blame stigma and generalization stigma seemed fuzzy, as comments often contained both types to a certain degree. Due to the limited context, comments were left to the most likely interpretation or were assigned to the unknown category.

In general, a higher annotation agreement rate was achieved for comments that show strong language with focus on blame, conflict, exaggeration, rejection, animosity, and aggression than comments with over-generalized statements - one-sided interpretation, unsupported judgement, and unfounded opinion. Moreover, stigmatised category is likely to be confused with hate speech; antagonism is part of the stigma concept but the latter encompasses a lot more than anger and hate sentiment. Anger can be a short emotional outburst or a response, whereas stigma is linked more to behaviour, acting on the emotion through rejection which can take several forms of antilocution, withdrawal, discrimination and eventually can result in very negative physical or continuous rejection (Allport et al., 1954).

Visualization results showed clear distinction between Stigma, Not Stigma and Undefined categories, which supports the goal of the study on achieving high inter-rate agreement and identifying characteristic features in each of the categories. Moreover anti-vaccination group emphasized blaming rhetoric towards government institutions and pharma companies. Where as pro-vaccination group places blame on anti-vaccination group, using expressions such as "*anti-vaxxer*", "*stupid*", "*want*", "*people*", "*die*".

As to prediction results, in Table 5 accuracy of each model is averaged over ten runs with its standard deviation and shows that most traditional models and deep learning models achieved accuracy over 75%. CNN significantly outperforms all other models according to the paired two-sampled t-test ($p < 0.05$) with 88% and 89% accuracy.

Table 5: Test accuracy on classification task. All models were evaluated ten times by boostrapping. We report mean and standard deviation of achieved accuracy. FastText (bigrams) outperforms baselines on the unbalanced datasets, whereas CNN significantly outperforms all other algorithms when evaluated on an upsampled, balanced dataset, as per a paired sample t-test ($p < 0.05$).

Model	Anti-Vaccination Balanced	Pro-Vaccination Balanced
TF-IDF, N-grams +Logistic Regression	0.821 ± 0.000	0.772 ± 0.000
+Support Vector Machine	0.839 ± 0.000	0.809 ± 0.000
+Naive Bayes	0.804 ± 0.000	0.770 ± 0.000
+MLP (Multilayer Perceptron)	0.858 ± 0.011	0.820 ± 0.007
+Random Forest	0.808 ± 0.000	0.752 ± 0.000
+K-Nearest Neighbours	0.754 ± 0.000	0.726 ± 0.000
+SGDC (Stochastic Gradient Descent)	0.837 ± 0.0039	0.812 ± 0.0046
LSTM	0584 ± 0.001	0.320 ± 0.016
BiLSTM	0.755 ± 0.044	0.769 ± 0.029
CNN	0.889 ± 0.010	0.885 ± 0.012
fastText 25 Epochs	0.582 ± 0.006	0.510 ± 0.004
fastText 25 Epochs, N-grams	0.592 ± 0.002	0.500 ± 0.006

Latter supports our goal on finding well performing deep learning model based on rigorous annotation and quality labelled data. Once the data has been balanced and increased by padding unbalanced classes with additional data points CNN improved its prediction accuracy as well.

6. Discussion

Relative scarcity of labelled data impedes use of deep learning models, however data scarcity is overcome by augmentation process as described in Zhong et al. (2017). Deep learning models and especially CNN show better performance accuracy after data augmentation with simple techniques such as padding, random erasing, cropping and flipping, etc. Data augmentation show advantage when deep learning models are applied on small classes, reduce the risk of over-fitting (Zhong et al., 2017),(Salamon and Bello, 2017),(Perez and Wang, 2017),(Han et al., 2018). Moreover, padded datasets are larger with higher number of features, where application of deep learning and CNN can achieve good performance due to its capacity to handle large number of features in comparison to other models. In Anti-vaccination discussions *not stigmatized* comments are high on emotional tone, are mainly informal, are expressed in shorter sentences, convey mainly *positive emotion*, and contain more *netspeak*, *assent* and *affect* words (for further elaboration on the type of words (see Table 7 in Appendix C). Stigmatised comments are expressed in lengthier sentences, use more formal language, are less emotional, contain more *negative connotation*, more *anger*, *power*, and *clout* words, make more references to *biological processes*, use less agreeable language (less *assent* words), and make more references to third person plural (e.g. "*they*"). Similarly, in pro-vaccination discussion on Facebook, stigmatised comments are expressed in lengthier sentences, show more *negative emotions*, *power*, *anger* words, and are less agreeable. In addition, more prominent features of health stigma in pro-vaccination discussions is a *lack of authenticity*, *swear words* and *death* references. Comments that do not express health stigma similarly to anti-vaccination context show *positive emotion*, *higher emotional tone*, are *more agreeable*, and contain *reward rhetoric*, *netspeak* and *informal words*. Not stigmatized comments in anti-vaccination context contain more references to *perceptual processed*, where perceptual emotion is based on capacities and abilities of recognizing and identifying emotions. In pro-vaccination discussions perceptual processes in not stigmatized comments are lacking, which can possibly mean that less empathy, fewer emotions and feelings are triggered in those discussions.

7. Conclusion and Future Work

In this work, the annotation approach was presented together with a discussion on challenges in designing an annotation scheme and annotating nontrivial labels. MTurk masters performed task with comparable or a higher accuracy than trained annotators, possibly due to their expert skill in performing similar tasks, whereas trained annotators never performed an annotation task before. Interesting findings in our study show that stigmatised language is expressed in less emotional but more formal ways, using *clout* and *power* features, whereas non-stigmatised language is less formal, has positive connotation, has more emotional features, and expresses sentiment using shorter words and sentences. Members of the pro-vaccination group express antagonism toward individuals who refuse to vaccinate (very often, these individuals are parents). On the other hand, members of anti-vaccination group use stigmatised rhetoric references of *them* - directed at government institutions, hospitals, or pharmaceutical companies. Differences in anger direction towards an individual in pro-vaccination group and anger directed at an institution in anti-vaccination group is due to the contrasting goals of the groups and might be seen as an obvious finding. However, this very finding serves as a confirmation of a proper dataset fit for the discourse analyses. Future work will be directed towards application of the annotation scheme to a different health-care domain and will include emotion lexicons from NLP community.

8. Bibliographical References

Allport, G. W., Clark, K., and Pettigrew, T. (1954). The nature of prejudice.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bello-Orgaz, G., Hernandez-Castro, J., and Camacho, D. (2017). Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems*, 66:125–136.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

De Brún, A., McCarthy, M., McKenzie, K., and McGloin, A. (2014). Weight stigma and narrative resistance evident in online discussions of obesity. *Appetite*, 72:73–81.

Du, J., Xu, J., Song, H.-Y., and Tao, C. (2017). Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with twitter data. *BMC medical informatics and decision making*, 17(2):69.

Faasse, K., Chatman, C. J., and Martin, L. R. (2016). A comparison of language use in pro-and anti-vaccination comments in response to a high profile facebook post. *Vaccine*, 34(47):5808–5814.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. *The handbook of social psychology*, 2(4):357–411.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Goffman, E. (2009). *Stigma: Notes on the management of spoiled identity*. Simon and Schuster.

Group, P. (2007). Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Han, D., Liu, Q., and Fan, W. (2018). A new image classification method using cnn transfer learning and web data augmentation. *Expert Systems with Applications*, 95:43–56.

Higuchi, K. (2016). Kh coder 3 reference manual. *Kioto (Japan): Ritsumeikan University*.

Jang, H., Piergallini, M., Wen, M., and Rose, C. (2014). Conversational metaphors in use: Exploring the contrast between technical and everyday notions of metaphor. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 1–10.

Joseph, A. J., Tandon, N., Yang, L. H., Duckworth, K., Torous, J., Seidman, L. J., and Keshavan, M. S. (2015). # schizophrenia: use and misuse on twitter. *Schizophrenia research*, 165(2-3):111–115.

Kang, G. J., Ewing-Nelson, S. R., Mackey, L., Schlitt, J. T., Marathe, A., Abbas, K. M., and Swarup, S. (2017). Semantic network analysis of vaccine sentiment in online social media. *Vaccine*, 35(29):3621–3638.

Katz, I. (2014). *Stigma: A social psychological analysis*. Psychology Press.

Kostkova, P., Mano, V., Larson, H. J., and Schulz, W. S. (2017). Who is spreading rumours about vaccines?: Influential user impact modelling in social networks. In *Proceedings of the 2017 international conference on digital health*, pages 48–52. ACM.

Li, A., Jiao, D., and Zhu, T. (2018). Detecting depression stigma on social media: A linguistic analysis. *Journal of affective disorders*, 232:358–362.

Link, B. G. and Phelan, J. C. (2001). Conceptualizing stigma. *Annual review of Sociology*, 27(1):363–385.

Massey, P. M., Leader, A., Yom-Tov, E., Budenz, A., Fisher, K., and Klassen, A. C. (2016). Applying multiple data collection tools to quantify human papillomavirus vaccine communication on twitter. *Journal of medical Internet research*, 18(12):e318.

Merriam-Webster. (2011). Dictionary by merriam-webster.

Mitra, T., Counts, S., and Pennebaker, J. W. (2016). Understanding anti-vaccination attitudes in social media. In *Tenth International AAAI Conference on Web and Social Media*.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

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

Saguy, A. C. (2012). *What's wrong with fat?* Oxford University Press.

Salamon, J. and Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283.

Schmidt, A. L., Zollo, F., Scala, A., Betsch, C., and Quattrociocchi, W. (2018). Polarization of the vaccination debate on facebook. *Vaccine*, 36(25):3606–3612.

Shutova, E. and Teufel, S. (2010). Metaphor corpus annotated for source-target domain mappings. In *LREC*, volume 2, pages 2–2.

Stuber, J., Meyer, I., and Link, B. (2008). Stigma, prejudice, discrimination and health. *Social science & medicine (1982)*, 67(3):351.

Tomeny, T. S., Vargo, C. J., and El-Toukhy, S. (2017). Geographic and demographic correlates of autism-related anti-vaccine beliefs on twitter, 2009-15. *Social Science & Medicine*, 191:168–175.

Wallington, A., Barnden, J., Buchlovsky, P., Fellows, L., and Glasbey, S. (2003). Metaphor annotation: A systematic study. *COGNITIVE SCIENCE RESEARCH PAPERS-UNIVERSITY OF BIRMINGHAM CSRP*, 2(2):3–4.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang,

Y. (2017). Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*.

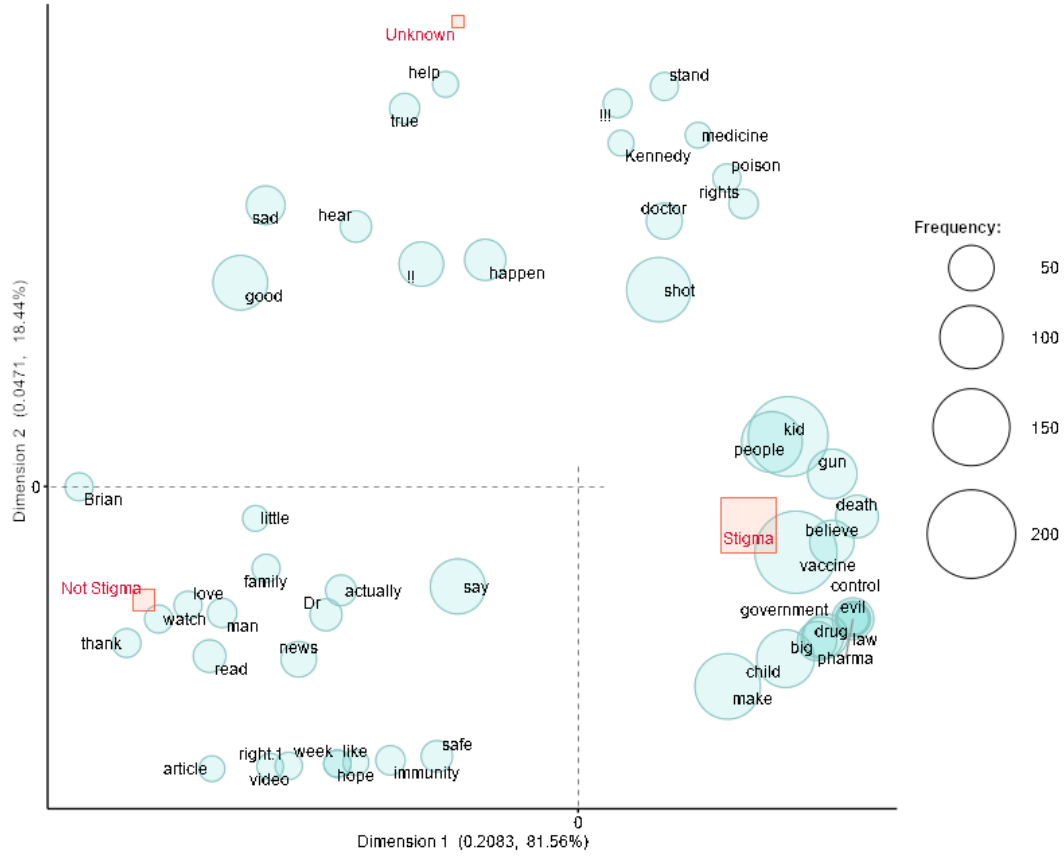
Appendix A

Table 6: Annotation scheme - 9 labels

Does the sentiment convey stereotype/prejudice/bias/stigma? If YES please choose an option from 1-8, if NO then 9 (NONE)	
Label	**Post/Comment Example**
1. Blame, Accusation, Judgement, Suspicion	".. NASA faked the moon landings do you really want to believe anything they say? NASA steals 50 million USD a DAY and all we get is really bad CGI and "space bubbles".
2. Conflict, Hate, Fear	"Boycott his movies", "He's a narcissistic coke head, This guy is a stick SOBb", "They are definitely spraying something."
3. Exaggeration, Strong emotion	"Antichrist", "Evil pure evil !"
4. Generalization	".. only Korea and Japan have good ones they are smart and know how to manage them but they have no illegals or refugees either those ruin healthcare they do not pay into it."
5. Predicting, guessing	"They may be injecting mothers before babies are born so they can say the baby was born with autism and it is genetic."
6. Personal opinion	"..I can guarantee the brainwashed will be flocking to get the jab! They've probably not made as much money on the flu jabs this year.. Scaremongering!"
7. "He/she/they" followed by "always, all the time, all"	"..No kidding, it's cuzz they don't protect you from anything, govt way to inject poisen called population control!"
8. I/us/we vs. them	", we don't have any rights, they bully us into it by slapping a mask on, when there's no medical or scientific evidence of the mask preventing it if you're asymptomatic, but the hospitals have to meet 90% compliance in order to get reimbursed by medicare thanks to our great health care system"
9. NONE of the above	"When mine were removed the doc did nothing safe. 15 + years ago", "Thank you Sentor Folmer. We are rooting for the change!", "I have 4 cats and three horses .. the amount of vaccines that they are required to get is staggering", "Let's keep our kids strong."

Appendix B

(a) Anti-Vaccination



(b) Pro-Vaccination

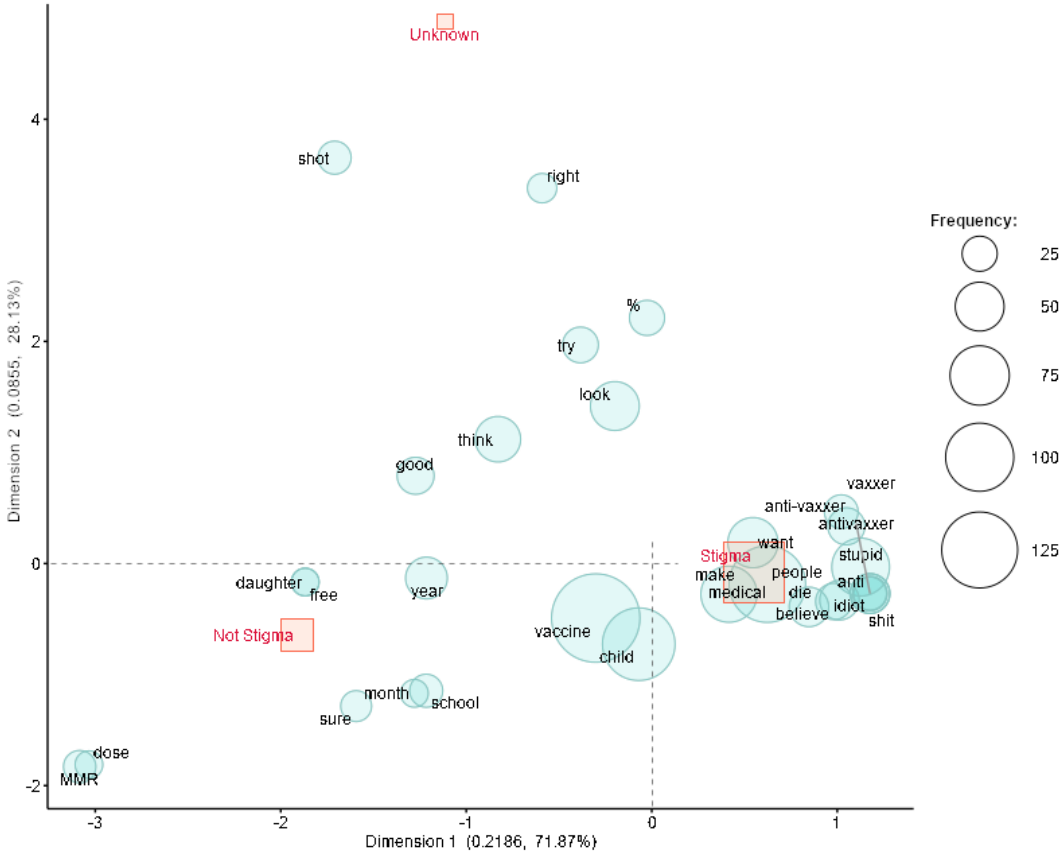


Figure 1: Correspondence analyses of words (see text for details).

Appendix C

Table 7: Selection of top 30 features with ANOVA F-value to measure the strength of correlation between labels and features). ANOVA F-value calculated based on 93 features.

Anti-Vaccination					**Pro-Vaccination**				
Features	**F-value classification score**	**Stigma**	**Undefined**	**Not Stigma**	**Features**	**F-value classification score**	**Stigma**	**Undefined**	**Not Stigma**
		Z-score: Std. Err. of the MEAN					**Z-score: Std. Err. of the MEAN**		
Words per Sentence (mean)	138.40	9.85	-7.16	-9.21	WPS (mean)	41.77	4.63	-7.14	-2.50
Word Count	86.38	8.04	-5.58	-7.68	WC	41.73	4.56	-7.23	-2.34
Positive emotion	63.09	-6.76	1.78	8.21	FocusPast	30.71	-3.79	-0.86	6.61
Emotional tone	51.17	-5.69	-0.25	7.96	Authentic	30.27	-4.01	-0.13	6.47
Assent	45.81	-4.77	-1.96	7.73	I	29.97	-3.82	-0.64	6.51
Informal words	39.62	-4.97	-0.44	7.09	Negative emotion	27.84	4.42	-3.01	-4.98
Article	28.37	4.51	-4.89	-3.24	Emotional Tone	27.30	-3.90	0.14	6.10
Prepositions	27.51	4.24	0.16	-5.92	Positive emotion	25.80	-4.01	0.97	5.72
Affect	24.56	-4.44	3.38	4.06	Anger	24.62	4.17	-2.83	-4.70
Reward	18.74	-3.89	3.01	3.53	Number	22.13	-3.42	-0.20	5.58
They	18.47	3.70	-0.66	-4.69	Netspeak	20.35	-3.71	3.93	3.22
Negative emotion	16.87	1.98	2.97	-4.50	All Punctuation	17.24	-2.66	5.10	0.76
Functional words	16.60	3.69	-2.16	-3.77	Biological processes	12.57	3.01	-2.43	-3.13
Exclamation mark	16.23	-3.60	3.00	3.14	Clout	11.20	2.80	-1.39	-3.50
Cause	11.72	3.10	-1.56	-3.32	Functional words	10.42	2.07	-3.99	-0.57
Anger	11.41	1.47	2.68	-3.63	Reward	10.34	-2.50	0.33	3.75
Perception	10.80	-2.34	-1.04	3.84	Dictionary words	10.10	0.78	-4.11	1.56
Quantifiers	10.43	2.25	1.13	-3.77	Social	9.07	2.49	-2.67	-2.15
Netspeak	9.15	-2.74	1.25	3.01	Swear words	9.04	2.28	-0.09	-3.56
Health	9.12	1.93	1.42	-3.51	Death rhetoric	8.77	2.51	-2.21	-2.49
Conjunction	9.02	2.74	-1.58	-2.81	Perception	8.76	-1.28	3.91	-0.63
Power	8.00	2.58	-1.39	-2.70	They	8.19	2.44	-1.69	-2.73
Clout	7.91	2.56	-1.33	-2.71	Relativity	7.44	-2.03	-0.05	3.26
Dictionary words	7.84	-2.01	-0.86	3.28	Words>six letters	7.37	2.11	-0.25	-3.18
Auxiliary Verb	6.91	2.39	-1.25	-2.53	Assent	7.33	-2.30	1.42	2.69
Period	6.82	-1.75	3.20	0.48	Question Mark	6.91	-1.68	3.27	0.45
Biological processes	6.69	2.21	-0.24	-2.89	Quantifiers	6.85	2.06	-2.69	-1.45
Question Mark	6.48	-1.35	3.32	-0.14	Article	6.72	2.00	-2.76	-1.30
See	6.27	-1.96	-0.37	2.91	Body references	6.39	2.01	-0.44	-2.90
Hear	6.15	-1.59	-1.16	2.89	Informal words	6.27	-1.91	2.71	1.20