Human Vaccines & Immunotherapeutics

ISSN: 2164-5515 (Print) 2164-554X (Online) Journal homepage: https://www.tandfonline.com/loi/khvi20

A natural language processing framework to analyse the opinions on HPV vaccination reflected in twitter over 10 years (2008 - 2017)

Xiao Luo, Gregory Zimet & Setu Shah

To cite this article: Xiao Luo, Gregory Zimet & Setu Shah (2019) A natural language processing framework to analyse the opinions on HPV vaccination reflected in twitter over 10 years (2008 - 2017), Human Vaccines & Immunotherapeutics, 15:7-8, 1496-1504, DOI: 10.1080/21645515.2019.1627821

To link to this article: https://doi.org/10.1080/21645515.2019.1627821

Accepted author version posted online: 13 Jun 2019.
Published online: 16 Jul 2019.

HUMAN VACCINES & IMMUNOTHERAPEUTICS
2019, VOL. 15, NOS. 7–8, 1496–1504
https://doi.org/10.1080/21645515.2019.1627821

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

A natural language processing framework to analyse the opinions on HPV vaccination reflected in twitter over 10 years (2008 - 2017)

Xiao Luo [a], Gregory Zimet [b], and Setu Shah [a]

[a]Purdue School of Engineering and Technology, Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA; [b]Pediatric Adolescent Health, Indiana University School of Medicine, Indianapolis, IN, USA

ABSTRACT

In this research, we developed a natural language processing (NLP) framework to investigate the opinions on HPV vaccination reflected on Twitter over a 10-year period – 2008–2017. The NLP framework includes sentiment analysis, entity analysis, and artificial intelligence (AI)-based phrase association mining. The sentiment analysis demonstrates the sentiment fluctuation over the past 10 years. The results show that there are more negative tweets in 2008 to 2011 and 2015 to 2016. The entity extraction and analysis help to identify the organization, geographical location and events entities associated with the negative and positive tweets. The results show that the organization entities such as FDA, CDC and Merck occur in both negative and positive tweets of almost every year, whereas the geographical location entities mentioned in both negative and positive tweets change from year to year. The reason is because of the specific events that happened in those different locations. The objective of the AI-based phrase association mining is to identify the main topics reflected in both negative and positive tweets and detailed tweet content. Through the phrase association mining, we found that the main negative topics on Twitter include "injuries", "deaths", "scandal", "safety concerns", and "adverse/side effects", whereas the main positive topics include "cervical cancers", "cervical screens", "prevents", and "vaccination campaigns". We believe the results of this research can help public health researchers better understand the nature of social media influence on HPV vaccination attitudes and to develop strategies to counter the proliferation of misinformation.

ARTICLE HISTORY
Received 23 March 2019
Revised 20 May 2019
Accepted 28 May 2019

KEYWORDS
Natural language processing; HPV vaccine; social media; public health; sextually transmitted infection

Introduction

Human papillomavirus (HPV) is the most common sexually transmitted infection (STI) in the United States and in the world. Persistent HPV infection can progress to several types of cancer, including cervical, anal, and oropharyngeal cancers. First licensed in the U.S. in 2006, HPV vaccine currently is approved for males and females ages 9 through 45 years for the prevention of cervical, anal, vaginal, and vulvar cancers, as well as genital warts.[1] Ninety-one countries had implemented national HPV vaccination programs as of June 2019, with most of these countries located in the Americas and western Europe. Relatively few countries in Asia have adopted national HPV vaccination programs.[2] Despite the recommendation and availability of safe and effective vaccines for over 10 years, HPV vaccination rates in the U.S. (and many other countries) are far lower than the goal set by Healthy People 2020 of 80% series completion for both adolescent males and females.[3] Identified reasons for non-vaccination include poor quality recommendations on the part of health care providers and unwarranted parental concerns about HPV vaccine safety and effectiveness. The extent to which parental concerns are influenced by social media stories has received quite a bit of attention.

Twitter, a social media platform created in 2006, has hundreds of millions of users around the world. It is a place for users to express their opinions and is frequently used for microblogging. The application's nature makes it an ideal medium for the study of online opinions and behaviours. Research has been done to analyse tweets in Twitter to understand the public concerns and influences on HPV vaccine uptake. Shapiro et al.[4] conducted a comparison of tweets that express concerns about HPV vaccines from three countries: Australia, Canada, and the UK. The study period was about two and a half years from January 2014 to April 2016. A model was created to distinguish between tweets with and without (neutral) concerns. The tweets expressing concerns about psychological barriers comprised the largest (46%) proportion of tweets expressing concerns. Dunn et al.[5] investigated HPV vaccine coverage in the United States by using tweets posted between October 2013 and October 2015. Their results indicated that the topics related to media controversies were most closely related to coverage (both positively and negatively), which explained variance in HPV Vaccine coverage in females in 2015. Le at al.[6] used three sampling strategies to analyse Tweets from 2014. The three sampling strategies were based on key words related to cervical cancer prevention. One hundred tweets from each of the three sampling strategies were used to examine the narratives and themes which were different across samples. Dunn et al.[7] analysed tweets related to HPV Vaccine for a six-month period (from Oct 2013 to April 2014). Machine learning models were built to classify tweets as anti-vaccine or otherwise, and their research identified that negative tweets contributed to 25.1% of

CONTACT Xiao Luo luo25@iupui.edu Purdue School of Engineering and Technology, Indiana University-Purdue University Indianapolis, 799 W. Michigan Street, Indianapolis, IN, USA

total HPV related tweets. Du et al.[8] extracted tweets between July 15, 2015 to August 17, 2015, and manually annotated 6000 tweets for sentiment analysis. Machine learning algorithm Support Vector Machine (SVM) was used to classify the tweets into 7 sentimental categories. The classification worked well on non-HPV related content and positive tweets, but not so well on the negative tweets.

In the present research, our main objective was to extend prior research efforts by developing a natural language processing framework to analyse HPV vaccine-related information extracted from one social media platform (Twitter) over a 10-year period (2008–2017). Natural language processing (NLP) is a subfield of computer science, computational linguistics, and artificial intelligence that helps computers process and analyse large amounts of natural language data. The NLP tasks include content categorization, topic discovery, entity recognition, sentiment analysis and so on. In this research, the NLP framework composites three main tasks organized as a pipeline to analyse the natural language data. Using NLP and artificial intelligence techniques, we sought to identify, and extract terms and phrases related to reasons for non-vaccination or opposition to HPV vaccination. These terms and phrases can help public health researchers identify the reasons for HPV vaccine scares and refine policies to prevent the scares and improve HPV vaccine uptake. Our research makes a unique contribution to the literature by evaluating opinions about HPV vaccination on Twitter over a 10-year timespan and by our development of an NLP framework to identify the associations between the phrases and entities reflected in both negative and positive tweets. Our intent in this paper is not to examine associations of Twitter trends with HPV vaccine policies or uptake, but to evaluate strengths and weaknesses of different Twitter analytic approaches. We hope to identify approaches that may prove valuable in helping researchers and public health professionals to understand events and associated phrases that might trigger the spread of negative opinion on social media. The results of a 10-year timespan provide a good overview and insight into attitudes about HPV vaccine reflected in social media.

Methods

Study data

English-language tweets (287,100 tweets) containing keywords related to HPV vaccines were collected between January 1, 2008 and Dec 31, 2017. These tweets were extracted by searching for any combination of keywords: HPV, vaccination, vaccine, cervical, cancer, and Gardasil, via a Twitter application programming interface (API), with respect to the terms of service for Twitter developers.

NLP analysis framework

In this research, we designed and developed an NLP framework by first classifying the HPV vaccine-related tweets into three groups: positive, negative and neutral through sentiment analysis. Then, to further understand the influential content mentioned in these groups of tweets, phrase association mining and entity analysis were applied to the positive and negative groups to understand the main opinion phrases or events that were associated to the positive and negative tweets. Figure 1 shows an overview of the framework. This framework can identify the negative tweets and the mentioned topics, events and entities within the tweets. Public health researchers may be able to make use of such a framework to identify negative opinions and associated public scares, public event and entities in the social media, so that preventive action can be taken to prevent drops in HPV vaccine uptake.

Sentiment analysis

Sentiment analysis inspects the given text and identifies the prevailing emotional opinion within the text to determine a writer's attitude as positive, negative, or neutral. The sentiment analysis used in this research was Google Cloud sentiment analysis. Google Cloud sentiment analysis belongs to the set of Cloud Natural Language API developed by Google.[9] Users need to obtain Google Cloud credentials and credits in order to use this Google Cloud service. Given a text, Google Cloud sentiment analysis produces sentiment scores and magnitude values for the scores. Google Cloud sentiment scores range from −1 to 1, where −1 is extremely negative, 0 is neutral, and 1 is extremely positive. Associated with each score is the magnitude value. A higher value of magnitude indicates the strength of the sentiment. Given a text that includes multiple sentences, the algorithm calculates the sentiment score and magnitude value for each sentence, and then provide an overall sentiment score and magnitude value for the input text. Figure 2 provides an example of the sentiment

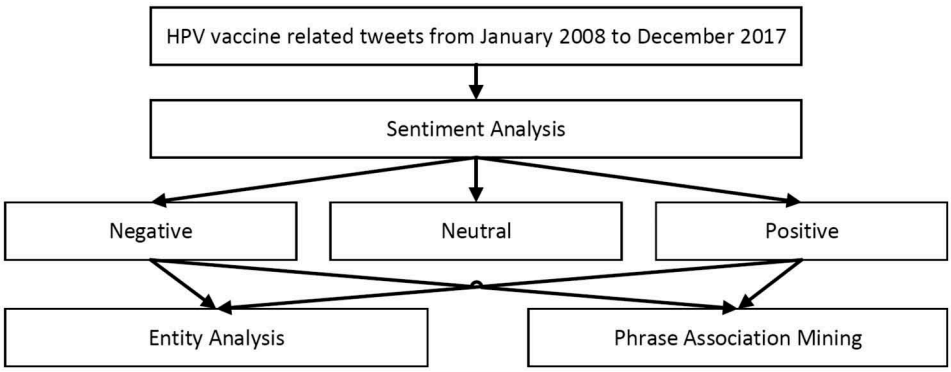

Figure 1. An NLP framework for HPV tweets analysis.

Figure 2. Example of sentiment analysis.

analysis. The sentiment score is 0.8, and the magnitude score is 0.8. That means this text has a very positive sentiment and an equally strong magnitude value.

The initial phrase cloud of positive tweets included phrases such as "scandal", "adverse event". Hence, those tweets were extracted out to be further analysed manually. We confirmed that these were negative tweets, demonstrating that Google Cloud Platform (GCP) sentiment analysis algorithm had some limitations. Similarly, we found that GCP sentiment analysis gave a sentiment score and a magnitude value of 0.9 and 0.9 respectively to the sentence, "Former Doctor Predicts that Gardasil will Become the Greatest Medical Scandal of All Time." However, the sentiment score and magnitude values for "scandal" are −0.6 and 1.6 respectively. It is likely that the inclusion of the word "greatest" led to the incorrect scoring of the sentence as a whole. Hence, we manually went through all the tweets that contained "scandal", "adverse", "safety" and "injury" to validate the sentiment analysis results and categorize them into the correct sentiment category. In total, we manually categorized 3672 tweets.

Entity extraction and analysis

The entity extraction and analysis approach involve more than simply scanning for nouns, but also inspects the given text for known entities such as public figures, landmarks, and common nouns such as restaurant, stadium, and so on. It also returns information about those entities. In the present study, spaCy entity analysis was used.[10] The spaCy's entity analysis identified the entities in text and classified them into many categories, some of which were: person, location, organization, event, and work of art. Figure 3 demonstrates an example of entity analysis by using spaCy's entity analysis. We used entity analysis to identify the critical entities mentioned in the positive and negative tweets of each year. We hypothesized that some entities, such as events or organizations might be associated with more negative tweets, and negative tweets might be linked to certain geographic locations.

Phrase association mining

The word embedding model proposed by Mikolov et al.[11] has been widely used for biomedical text analysis.[12] One advantage of word embeddings is that it discovers the semantic associations between words based on the content of the text. We believe that opinions and entities related to the opinion are often represented as phrases or terms which contain one or more than one word. Hence, in this research, we made use of phrase embeddings[11] to discover the associations between phrases within the tweets. First, the most common phrases in the text corpora were identified and tagged in the corpora with phrases instead of words. Lastly, the word embedding model was employed with the newly found phrases. In this study, we employed the Skip-Gram model. The neural network architecture of the Skip-Gram is a standard three-layer neural network. The inputs to the neural network are phrases that are represented as vectors (V_c). The length of the vector is the number of phrases (N) in the corpora. The output vectors ($U_{c-m} \cdots U_{c+m}$) are those to be evaluated against the defined context phrases of an input phrase. The context phrases are the ones that occur within a specific sliding window ($2m$) of the input phrase in a sentence. The training objective is to minimize the cost Jwhich is presented as Equation 1:

$$J = -\sum_{j=0.\neq m}^{2m} u_{c-m+j}^{T} v_c + 2mlog \sum_{k=1}^{N} \exp\left(u_k^T v_c\right) \quad (1)$$

At the end of the training process, each phrase is represented by a vector, the association scores between the phrases x and y can be calculated through a distance measure. In this study, the cosine distance (cos), as presented as Equation 2, was used to calculate the association scores. In this research, we made use of the phrase association mining to understand the phrases that were mostly related to the keywords of HPV vaccines in the positive and negative tweets respectively.

$$\cos\left(v_x, v_y\right) = \frac{v_x \cdot v_y}{\|v_x\| \|v_y\|} \quad (2)$$

Results

The number of tweets from each year is listed in Table 1. The number of tweets increased sharply from 2008 to 2009, peaked in 2013, but has remained over 30,000 per year since 2013.

Sentiment analysis

Sentiment analysis was done on the tweets of every year. The time series boxplot in Figure 4 shows the progress of public opinion on HPV vaccine over time, illustrated by the boxes that show proportionally more negative tweets in from 2008 to 2011, then, equivalent amounts of negative and positive tweets from 2012 to 2014. Finally, negative tweets

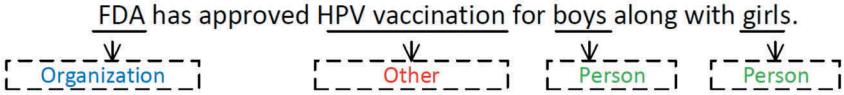

Figure 3. Example of entity extraction and analysis.

Table 1. Number of tweets of each year.

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Number	581	11372	14895	41571	27865	48203	33278	40898	35600	32837

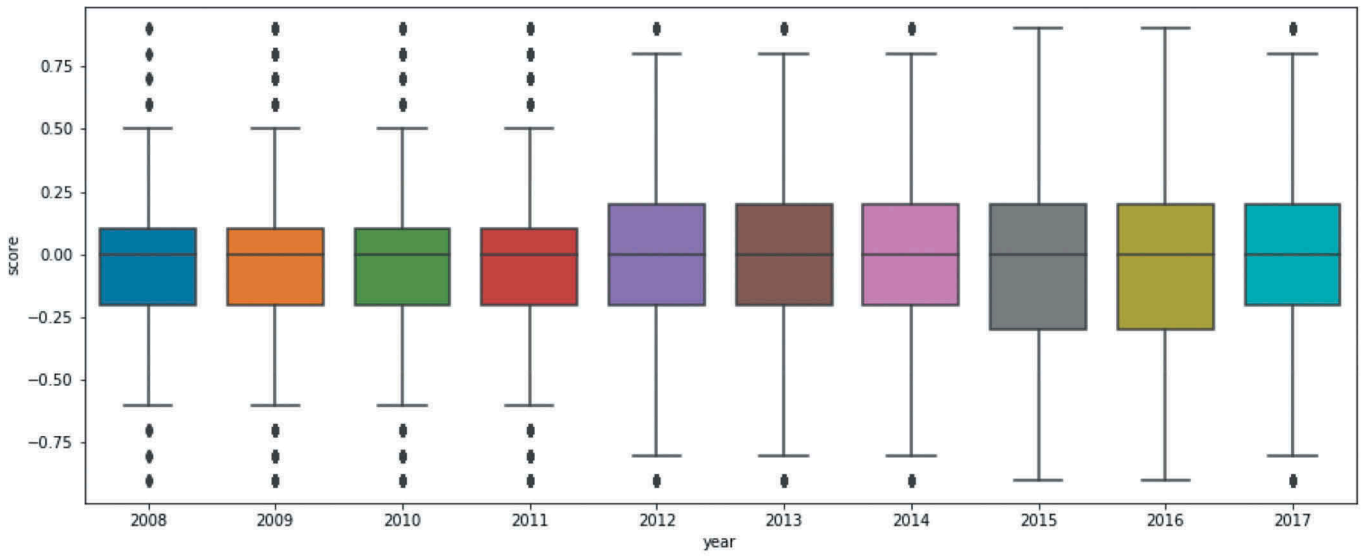

Figure 4. Boxplot of sentiment scores changes over 10 years.

predominated again in years 2015 and 2016. Word cloud is a visualization method for text which is straightforward and appealing. It has been used by many studies in text mining and opinion mining.[13] In our study, the data set had a large amount of text data about people's attitudes toward the HPV vaccine. The phrase cloud can provide a general overview of phrases that were used to express opinions over the past 10 years. Figure 5(a,b) provides phrase clouds of positive and negative tweets respectively. The phrase cloud of the positive tweets shows 'cervical cancer', while this term does not appear in the phrase cloud of the negative tweets. Similarly, the word "clean" is prominent in positive tweets, but not in negative tweets. Conversely, the word "scandal" is prominent in negative tweets, but not positive tweets. Interestingly, several words appear in both word clouds (e.g., "HPV Vaccine" and "Gardasil Vaccine"), indicating some limitations to this approach. It emphasizes the phrases based on the frequency in the study data.

Entity extraction and analysis

Based on the results of the sentiment analysis, entity extraction and analysis were applied to the positive and negative tweets of each year. We found that "CDC" is the organization entity that occurred in both positive and negative tweets for almost all the 10 years, except 2010. Other than 2008, 2009 and 2016, "CDC" occurred more in the negative tweets than positive tweets. After further investigation, we found that most of the negative tweets mentioned "CDC" because either they tagged "CDC" or wanted "CDC" to stop supporting the vaccine. Similar patterns were apparent with "FDA" and "Merck". Figure 6 shows the frequency of "CDC", "FDA" and "Merck" in the positive and negative tweets over the 10 years. Table 2 shows examples of tweets associated with organizations like CDC, FDA, and Merck for some of the years. We found that the high frequencies of some tweets were due to repeated postings. For example, the tweet, "Gardasil (MERCK) Vaccine against HPV human papilloma virus Sale and personal delivery throughout the country", was repeated over one thousand times in 2013, which caused the spike of "MERCK" in the positive tweets of that year.

Starting from 2010, many positive and negative tweets had location entities in them. In 2013, 'Southern US' was one of the popular location entities in the negative tweets. We found that this was because of the tweet – "HPV vaccination rates alarmingly low among young women in Southern US", which actually was not reflective of a negative sentiment about the HPV vaccine. In 2015, 'Central and South America' was a frequent location entity. We found that it was linked to the tweet, "HPV Vaccine Injuries and Deaths Now Being Reported from Central and South America". Tables 3 and 4 detail the most frequent location entities in the positive and negative tweets along with detailed tweets.

The event entity extraction and analysis on the negative and positive tweets showed that there was no event that was frequently associated with HPV vaccine-related tweets for a -specific year. One exception was in 2016, when "World Cancer Day 2016" was mentioned in 7 tweets. These 7 tweets were about "HPV news: On World Cancer Day 2016: Study explains why HPV vaccination rates are low in Hawaii", which were categorized as neutral to positive tweets.

Phrase association mining

Phrase association mining was used to identify the most relevant phrases to the keywords of the HPV vaccine and the key reasons extracted from the NIS-teen survey[14] for non-

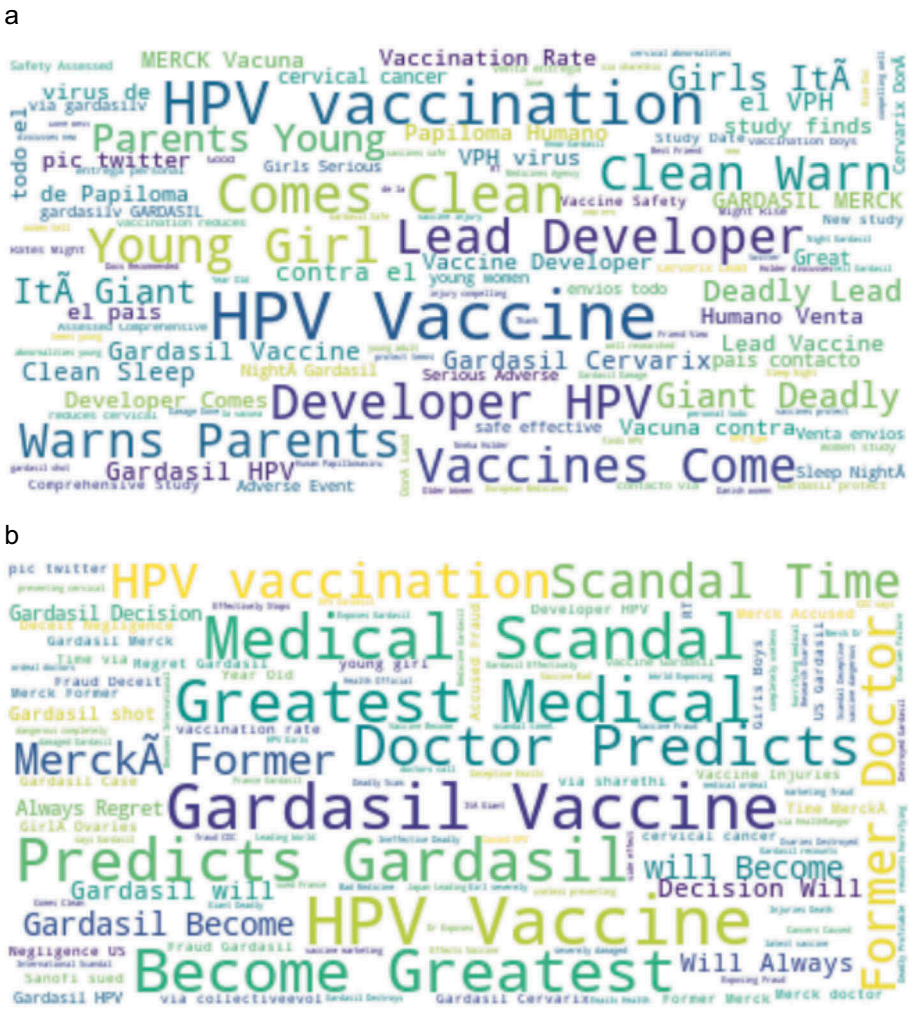

Figure 5. (a) Phrase cloud of positive tweets. (b) Phrase cloud of negative tweets.

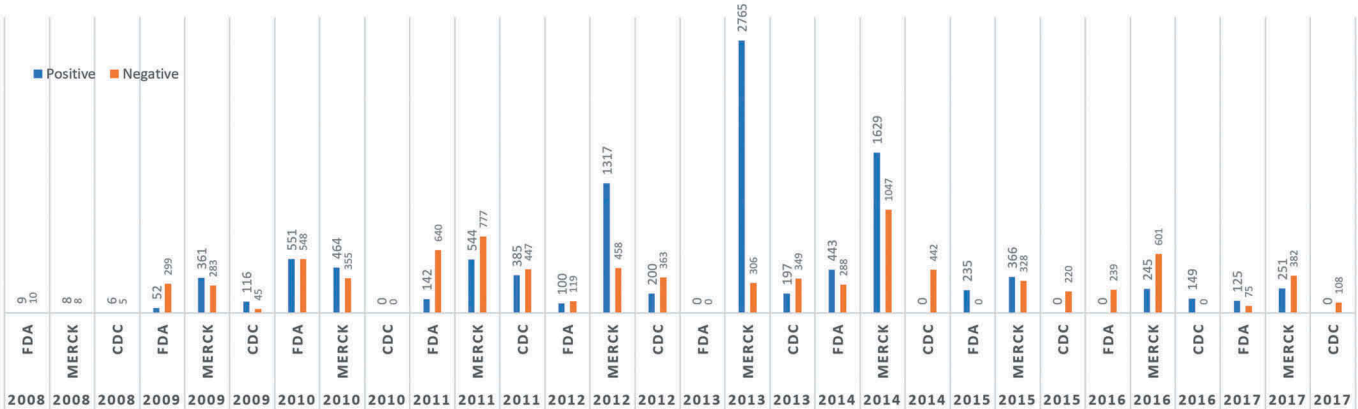

Figure 6. Frequency of "CDC", "FDA" and "Merck" in the positive and negative tweets.

vaccination. The phrase association mining algorithm was employed on the negative and positive tweets of each year to investigate the changes in the phrase associations. In this paper, we examined the identified negative phrases/words, their association scores with the keywords of the HPV vaccine and the content that reflected the associations. Based on the phrase cloud of the negative tweets, the negative phrases were those containing the following words: 'death', 'concern', 'kill', 'injured', 'safety', 'adverse', 'scandal' and 'fraud'. Table 5 shows selected phrase associations for each year and the changes over the past 10 years. It shows that in the early years before 2010, the HPV vaccine was highly associated

Table 2. Sample positive and negative tweets contain "CDC", "FDA", or "MERCK".

Entity	Sentiment	Sample Tweets (year)
"CDC"	Positive	CDC Panel Recommends HPV Vaccine Cervarix For Girls, Optional Gardasil Vaccination For Boys (2009) The CDC considers Gardasil safe and effective and says its side effect profile is comparable to other vaccines. (2013) RT opa1: HPV vaccines protect against types of HPV that cause cervical cancer. CDC recommends that girls & boys aged 11–12 get three … (2013) Via @NPR: CDC Endorses A More Effective #HPV Vaccine To Prevent Cancer (2016)
	Negative	10–21-10 VAERS Data for HPV Vaccines. How many deaths will it take for the CDC and FDA to take notice. (2010) A Plea to Merck, FDA & CDC to take action & stop all Gardasil vaccination. The evidence of harm caused by Gardasil … (2011) Updated: In latest vaccine marketing fraud, CDC says Gardasil shots should be 'routine' for boys (2012) #Gardasil kills more than the thing its suppose to protect us against. #fda #cdc why is it still on the market??? (2013) I advised against the Gardasil vaccine 3 yrs ago. NOW the CDC says OOPS only 2, NOT 3 shots needed: Maybe need 0. (2017)
"FDA"	Positive	FDA approves Gardasil for boys, young men (2009) FDA approves Gardasil 9 for prevention of certain cancers caused by five additional types of HPV: (2014) It's CervicalCancer Awareness Month. Learn about 3 US FDA approved HPV vaccines that can prevent it. (2016) Great news that the China FDA has approved two types of HPV vaccines in mainland China. I imagine it is still a bit cost–prohibitive for many. (2017)
	Negative	New Blog post FDA Again Says No to Expanded Gardasil Vaccine Use … (2009) FDA Records Detailing 16 New Deaths Tied 2 Gardasil: serious reports, w/213 resulting in permanent disability. (2010) FDA approves Gardasil vaccine: Gardasil will not prevent the development of anal precancerous lesions associated … (2011) HEALTH ALERT! #Merck won't explain why 100s of #girls are having #seizures months/yrs after #Gardasil #HPV #vaccine #FDA #CDC (2012) Israel Health Ministry Considers Cancelling HPV Vaccination Due to Side Effects … Israeli scientists, had warned the FDA & CDC about the … (2013) VAERS shows almost 1000 Disabled from Gardasil that is with 1–10% reporting, wake up America our kids are sick FDA does nothing (2013) @SaneVaxx 40 young women died in the trials of #HPV #vaccines it should have been banned according to #FDA protocols. (2017)
"MERCK"	Positive	Canada Drugstore FDA Approves GARDASIL(R) For Use In Boys And Young Men: Merck & Co., Inc. ann. (2009) RT fwpharma: Two doses of Merck & Co.'s Gardasil as effective as three doses in young girls, study finds (2013) Top 5 Vaccines by 2020–1) Prevnar pfizer 2) Pentacel sanofi 3) Gardasil Merck 4) Fluzone sanofi 5) Pediarix GSK (2014) Merck Systematic Review of 58 Publications of Real-World Use of GARDASIL Presented at EUROG … Read more: (2016)
	Negative	U.S. denies approval to expand use of Gardasil vaccine: The U.S. Food and Drug Administration has asked Merck & . (2009) New blog post: Merck Researcher Admits: Gardasil Guards Against Almost Nothing (2010) Not only were the Gardasil manufacturers Merck and other promoters unaware of the presence of dangerous … (2012) US court pays $6 million to Gardasil victims: Sorry Merck You Bastards Got Caught, "It's Not Safe" (2013) PARENTS Question the Vaccine! Was Merck Deceitful about Research Concerning Ovarian Failure and GARDASIL VACCINE? (2017)

Table 3. Top frequent location entities in positive tweets.

Year	Location (Frequency)	Sample Tweets
2010	Island (4)	HPV vaccination high among Maori and Island women: Young Maori and Pacific Island women are leading the way in the … HPV vaccination high among Maori and Island women: … Island women are leading the way in the uptake of the HPV
2011	Europe (18)	Europe authorizes the use of an HPV vaccine in children: Gardasil can be administered to children from … Europe authorizes use of papillomavirus vaccine in children and young people. "Gardasil" Prevents diseases linked to HPV and its propagation Europe approves papilloma vaccination for men with Gardasil. What good! Now, to see if we are financed like women.
2012	Europe (14)	HPV vaccine now use to protect schoolgirls – Pharmacy Europe: … EU advises all girls need HPV vaccines: All girls in Europe should be immunised against the … ECDC updated guidance on HPV vaccination in Europe: Focus on reaching all girls
2013	Africa (11)	$5/dose Gardasil for 8 countries in Asia/Africa, to vaccinate abt 180K girls GAVIAlliance could you talk about GAVI's efforts to expand HPV vaccines to southern Africa to prevent cervical cancer? "This is an important moment for Kenyan women" HPV Vaccines reach #Kenya! Important step for #womenshealth in #Africa
2014	Europe (95)	Europe approves the indication of 'Gardasil' (Sanofi Pasteur MSD) for the prevention of anal cancer RT WHO Europe VPI: Ireland achieved 81.9% uptake among 12–yr–old girls in 1st yr of HPV vaccination. Gardasil: New 2–dose Schedule Granted Positive CHMP Opinion for Europe's Leading HPV Vaccine The European Union gives the green light to the indication of the Gardasil vaccine against HPV for the prevention of cancer …
2015	Europe (100)	Vaccine. Organization and quality of HPV vaccination programs in Europe. May 14, 2015 More than 200 doctors across Europe are calling for a halt on the routine use of the HPV vaccination Europe recommends Gardasil 9, a new HPV vaccine that covers 5 additional antigens for immunization HPV vaccination for boys gets strong backing at European Cancer Conference today ECC2015
2016	Northern Cape (11)	Northern Cape wakes up to HPV vaccination for girls Northern Cape Resumes HPV Vaccination
2017	Europe (6)	Breakthroughs like HPV vaccination help to prevent cervical cancer affecting more than 58,000 women in Europe every year WHO Europe resource Talking with patients & parents about HPV vaccination for girls

with 'safety concerns' and 'death'. Starting from 2011, more negative tweets were about 'injuries', 'fraud' and 'scandal'. The 'adverse reaction/effects' and 'side effects' were the concerns from 2011 until 2017. Table 6 shows selected phrase association identified from the positive tweets. It consistently shows that the positive tweets were often associated with `cervical cancer' and/or `prevention'.

Discussion

In this descriptive study of HPV vaccination on Twitter, the overall results showed that many concerns about HPV vaccine could be discovered from using such a multi-faceted approach

Table 4. Top frequent location entities in negative tweets.

Year	Location (Frequency)	Sample Tweets
2010	Europe (2)	@SonyaVanSickle My mistake. Gardasil is still sold in Europe
		@SonyaVanSickle AGREED! What are people thinking with Gardasil? From my understanding Europe has even dropped the drug as too dangerous.
2011	America (7)	RT Americas Hope Gardasil boys come
2012	Europe (3)	marching in – 4 new Gardasil deaths reported
2013	Southern US (14)	to VAERS
		PP/Vaccine shills now pushing Gardasil jab on teens who don't even have sex. Europe Requires Warning Labels on Foods with Artificial Colors
		HPV vaccination rates alarmingly low among young women in Southern US Why is the HPV vaccination less likely to happen in the south?
		HPV vaccination less likely to happen in South than elsewhere: Initiation and completion rates for the human *p* … Be careful in the South!
		HPV vaccination rates alarmingly low among young women in Southern US
2015	Central and South America (23)	HPV Vaccines: Updates from Central and South America. First Reported Fatality after Gardasil in Colombia
		HPV Vaccine Injuries and Deaths Now Being Reported from Central and South America. HPV Gardasil vaccines
2017	Europe (5)	Gardasil is being banned in Europe because of the high rate of vaccine related injury
		China and Japan both stopped gardasil cause of severe reactions.Europe does not have the same as Us schedule. we push double the vaccines

to analysis of Twitter posts over the past 10 years. The developed NLP framework has the capability to first analyze the sentiment of the tweets, then extract key phrases and analyze the association of the phrases to further understand the main topics of the negative tweets. The identified topic may help public health researchers identify the reasons for HPV non-vaccination. The developed NLP framework also has the capability to find location, organization and event entities within the tweets. This can enable public health researchers of different countries or states to compare and investigate how to improve HPV vaccination at different locations and through different organizations and events.

The number of tweets peaked in 2013. We hypothesized that the main reason was that awareness of HPV vaccine increased. Although we also found a voluntary recall of one lot of Gardasil HPV Vaccine issued by CDC in 2013,[15] we thought it was not the main cause for the peak, since the word "recall" was not one of the frequent phrases associated with negative tweets. Through phrase association mining, we identified some negative topics that were different from those most cited in the literature[16-19] as drivers of non-vaccination, such as "marketing fraud", "scandal", and "injury". These topics could cause HPV vaccine scares. Public health researchers could take appropriate action on providing clarifications on these topics and prevent scares, so that HPV uptake does not diminish. Through the entity analysis, we discovered that "CDC" was also associated with negative tweets. Some of the positive tweets were linked to CDC recommendation on HPV vaccine, whereas the negative tweets were initiated by Twitter users who expressed that CDC should take some action to stop the HPV recommendation. And we identified that Europe is the location entity that was mentioned almost every year. Although some of those tweets are positive, some are negative, reflecting variations in HPV vaccine hesitancy in European countries and the fact that HPV vaccination programs around the world have been introduced at different times.[20]

There are several limitations of the work reported here, which also serve as a cautionary tale for analysis of social media posts. One serious issue was that the GCP sentiment

Table 5. Phrase association mining for negative tweets.

Year	Keyword	Associated Phrase	Association Score	Tweets
2008	immunization	kill	0.905	Gardasil HPV vaccine for girls kills teens; do a 'google' or 'yahoo' search;
2009	Gardasilbeware	danger	0.543	The Latest News About Gardasil Vaccine Danger From Gardasilbeware
	raise concerns	safety continues	0.948	Gardasil safety continues to raise concerns
	Gardasil	deaths	0.871	28 Deaths Tied To Gardasil, Watchdog Says -
2010	vaccine	people concerned	0.543	Gardasil vaccine for cervical cancer unproven and dangerous.
	vaccine	adverse reactions	0.264	Gardasil HPV vaccine touted to fight cervical cancer, has killed 71 girls and caused thousands of adverse reactions.
	Gardasil	death	0.854	Vaccines-Cervarix, Gardasil, Death, And Extreme Side Effects
2011	gardasil	injured	0.521	1000s Injured gardasil learn more truthaboutgardasil.org kids are DYING Merck do something FDAmedwatch cdc health girls boys cancer
2012	hpv vaccines	injured	0.348	FDArecalls how about recalling Gardasil, 25,548 have been injured and 111 are DEAD! woman boys HPV Vaccines CDC Merck health girls
	vaccine marketing	marketing fraud	0.515	In latest vaccine marketing fraud, CDC says Gardasil shots should be 'routine' for boys
	Gardasil	adverse reaction	0.519	there are so many adverse reaction to Gardasil but the media does not publish them.
2013	cervical cancer	serious adverse	0.371	Gardasil has been associated with as many serious adverse events as there are deaths from cervical cancer developing per year
	Gardasil	fraudulent	0.588	Most Gardasil 'safety' data discovered to be fraudulent HPV vaccine cervarix
2014	vaccines	deaths	0.272	Supreme Court in #India to Rule on #Merck Fraud Regarding #HPV #Vaccine Deaths
2015	Gardasil	scandal	0.311	6248 Permanent Injuries and 144 Deaths Following Gardasil HPV Vaccine: Coincidence or Scandal?
2016	faulty vaccines	injuries caused	0.531	Irish teens taking on Gardasil manufacturers after suffering from long-term injuries caused by the faulty vaccines
	HPV vaccine	killed	0.312	Gardasil exposed: HPV vaccine is being tested on infants it has killed permanently injured thousand
	HPV vaccines	adverse effects	0.871	HPV Vaccines account for over 44% of all adverse reactions.
2017	HPV vaccines	side effects	0.363	the side effects for the hpv vaccines are so annoying i do not recommend

Table 6. Phrase association mining for positive tweets.

Year	Keyword	Associated Phrase	Association Score	Tweets
2008	cervical cancer	prevents	0.811	@violetblue I have one (of three) Gardasil shot to go, nice to know it could possibly prevent other nasties other than cervical cancer!
2009	continued vaccination	agency recommends	0.958	Published a new post: European Medicines Agency Recommends Continued Vaccination With Gardasil
	cervical disease	at preventing	0.642	Buy Vemma Merck's Gardasil Effective At Preventing HPV, Cervical Disease In Older Women, Study Finds: M.
2010	cervical cancers	included in	0.519	Two human papillomavirus types included in HPV vaccines cause 71% of cervical cancers
2011	cervical legions	evidence summary	0.630	Immunization News! Evidence summary shows HPV vaccination is helping prevent cervical legions in young women
2012	vaccination campaigns	best practices	0.803	MT Video webinars on best practices from HPV vaccination campaigns in LMIC countries from Cervical Cancer
	cervical cancers	cause most	0.793	Two vaccines, Cervarix and Gardasil, protect against most of the types of HPV that cause most cervical cancers
2013	vaccinations valuable	anal	0.789	hpv vaccinations valuable protection for boys too – protects against anal, penile, neck cancers
	vaccination rates	hpv vaccination	0.405	Great project in Kentucky to try and boost HPV vaccination rates!
2014	reduces cervical	lesion	0.953	HPV Vaccination Reduces Cervical Lesion Risk at Population Level: Cohort study in Denmark shows reduction in r …
2015	newer vaccinations	over time	0.942	@AJemaineClement @djnrrd We need unbiased, longer studies, and monitor over time. Especially on the newer vaccinations like Gardasil.
2016	cervical cancers	whole range	0.404	Ladies, protect yourselves against a whole range of cervical cancers by starting on an HPV vaccination course today!
2017	cervical screens	lifetime	0.670	HPV vaccination means women only need three cervical screens in a lifetime, study suggests …

analysis sometimes misidentified the nature of the sentiment expressed. This may have occurred because GCP sentiment analysis is not customized to evaluate sentiments that show consumers' resistance to, or opinions about, a medical product. Often, an opinion about a medical product is not simply expressed through use of terms such as "good", "bad", and "great". Although we manually corrected the sentiment analysis results of some tweets, because of the size of the corpus, it was not feasible to manually validate all of them. In the future, we plan to develop a semi-supervised learning algorithm for sentiment analysis on this study corpus. The semi-supervised learning algorithm will make use of the Long Short Term Memory (LSTM) Neural Networks[21] to classify the tweets to the sentiment class: positive and negative and make use of the concept embedding techniques[22] to consider the semantics of the words and phrases in the tweets. We also plan to annotate some categories, such as "fraud", "adverse", "side effect", "death" and so on for negative tweets based on the identified key phrases from the tweets, and then train the learning model to automatically identify the tweets in those categories. In this work, we simply classified the tweets into positive and negative based on threshold 0. More sophisticated thresholds can be used to identify extreme negative (<-0.75) and extreme positive (>0.75) tweets and the content within them. More tweet features, such as number of retweets, favorites, and replies might be used to identify the spreading of the negative and positive tweets on the social media.

In this largely descriptive study, we did not generally attempt to link specific historical events related to HPV vaccination to patterns of opinions reflected on Twitter over time. As we work to develop more accurate algorithms for sentiment analysis, it will be important to evaluate ways to use this information to quickly assess the impact of historical events, such as false reports of serious adverse events, on social media trends. In this way, it may be possible to intervene to limit the damage that can result from the rapid dissemination of false information on social media.

Conclusions

In this descriptive research, we explored an NLP framework which included sentiment analysis, entity analysis and phrase association mining to analyse tweets from social media about HPV vaccination. This framework can help analyse and identify the main topics and entities related to the positive and negative text buried in a large text corpus. Our results show the different negative and positive opinions on HPV vaccination reflected in Twitter over the past 10 years. Tweets about HPV vaccine significantly increased after 2010 and peaked in 2013. In earlier years, soon after the FDA licensed HPV vaccine, the main concern was 'death' and 'safety', later it was more about 'injuries', 'adverse effects' and 'fraud'. The results provide an overall insight into the concerns expressed through social media, which can help health authorities form and evaluate was of countering the misinformation and fear mongering that has been relatively common on Twitter and other social media platforms. At the same time, it is encouraging that a substantial proportion of tweets had neutral or positive sentiments associated with them. Future work in this domain includes adding emoji analysis into the sentiment analysis, given that users express their feelings and opinions not only through words but also through emojis; customizing the sentiment analysis algorithm to work better with tweets and to tailor it towards understanding the context of opinions in the medical domain. In addition, it will be important to

examine the extent to which negative or positive tweets influence actual uptake of HPV vaccine.

Disclosure of potential conflicts of interest

Within the last year, author Gregory Zimet received an honorarium from Sanofi Pasteur for work on the Adolescent Immunization Initiative and received travel support from Merck to attend a conference on HPV vaccination.

ORCID

Xiao Luo http://orcid.org/0000-0002-3649-9785
Gregory Zimet http://orcid.org/0000-0003-3835-937X
Setu Shah http://orcid.org/0000-0003-2951-2272

References

1. FDA. Retrieved from FDA approves expanded use of Gardasil 9 to include individuals 27 through 45 years old; 2018 Oct 5. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm622715.htm.
2. CervicalCancerAction, 2. Cervical cancer action. Global Maps: global Progress in HPV Vaccination: status: June 2017. 2017. www.cervicalcanceraction.org: http://www.cervicalcanceraction.org/comments/comments3.php.
3. Healthy People 2020, Immunization and infectious Disease. 2015 [accessed February 19 2019]. healthypeople.gov: https://www.healthypeople.gov/2020/topics-objectives/topic/immunization-and-infectious-diseases.
4. Shapiro GK, Surian D, Dunn AG, Perry R, Kelaher M. Comparing human papillomavirus vaccine concerns on twitter: a cross-sectional study of users in Australia, Canada and the UK. BMJ Open. 2017;7:e016869. doi:10.1136/bmjopen-2017-016869.
5. Dunn AG, Surian D, Leask J, Dey A, Mandl KD, Coiera E. Mapping information exposure on social media to explain differences in hpv vaccine coverage in the United States. Vaccine. 2017;35:3033–40. doi:10.1016/j.vaccine.2017.04.060.
6. Le GM, Radcliffe K, Lyles C, Lyson HC, Wallace B, Sawaya G, Pasick R, Centola D, Sarkar U. Perceptions of cervical cancer prevention on Twitter uncovered by different sampling strategies. Plos One. 2019;14:e0211931.
7. Dunn AG, Leask J, Zhou X, Mandl KD, Coiera E. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. J Med Internet Res. 2015;17:e144. doi:10.2196/jmir.4343.
8. Du JX, Xu J, Song H, Liu X, Tao C. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. J Biomed Semant. 2017;8. doi:10.1186/s13326-017-0120-6.
9. GCP. Natural language - Google cloud. Google Cloud; 2019. https://cloud.google.com/natural-language/.
10. Honnibal M, Montani I. spaCy - Industrial-Strength Natural Language Processing, 2019. https://spacy.io/
11. Mikolov T, Sutskever I. Distributed representations of words and phrases and their compositionality. Proceedings of the International Conference on Neural Information Processing Systems; Nevada, USA: Lake Tahoe; 2013. p. 3111–3119.
12. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, Kingsbury P, Liu H. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inf. 2018;87:12–20. doi:10.1016/j.jbi.2018.09.008.
13. Lemire OK. Tag-cloud drawing: algorithms for cloud visualization. The Proceedings of International World Wide Web Conferences; Banff (Canada): ACM; 2007. doi:10.1094/PDIS-91-4-0467B.
14. National Immunization Survey-Teen (NIS-Teen). National Immunization Surveys (NIS), CDC; 2017. https://www.cdc.gov/vaccines/imz-managers/nis/datasets.html
15. 2013 voluntary recall of one lot of Gardasil HPV vaccine. Center for Disease Control and Prevention; 2013. https://www.cdc.gov/vaccinesafety/concerns/history/gardasil-recall-faq.html
16. Kester LM, Zimet GD, Fortenberry JD, Kahn JA, Shew ML. A national study of HPV vaccination of adolescent girls: rates, predictors, and reasons for non-vaccination. Maternal Child Health J. 2013;17:879–85. doi:10.1007/s10995-012-1066-z.
17. Zimet GD, Weiss TW, Rosenthal SL, Good MB, Vichnin MD. Reasons for non-vaccination against HPV and future vaccination intentions among 19–26 year-old women. BMC Womens Health. 2010;10. doi:10.1186/1472-6874-10-27.
18. Holman DM, Benard V, Roland KB, Watson M, Liddon N, Stokley S. Barriers to human papillomavirus vaccination among US adolescents: a systematic review of the literature. JAMA Pediatr. 2014;168:76–82. doi:10.1001/jamapediatrics.2013.2752.
19. Why some parents are refusing HPV vaccine for their children. Retrieved from shot of prevention. 2013. https://shotofprevention.com/2013/08/20/why-some-parents-are-refusing-hpv-vaccine-for-their-children/
20. Karafillakis E, Simas C, Jarrett C, Verger P, Peretti-Watel P, Dib F, De Angelis S, Takacs J, Ali KA, Pastore Celentano L, et al. HPV vaccination in a context of public mistrust and uncertainty: A systematic literature review of determinants of HPV vaccine hesitancy in Europe. Human Vaccine Immunother. 2019;1–13. doi:10.1080/21645515.2018.1564436.
21. Sundermeyer MS. LSTM neural networks for language modeling. In Thirteenth annual conference of the international speech communication association; 2012. doi:10.1094/PDIS-11-11-0999-PDN.
22. Daniel C, Yinfei Y, Sheng-yi K, Nan H, Nicole L, Rhomni SJ, Noah C, Mario GC, Steve Y, Chris T, et al. Universal sentence encoder. 2018. arXiv preprint arXiv:1803.11175.