# Supporting users with credibility assessments of health Tweets

*Master's Thesis*

Freek Post

# Supporting users with credibility assessments of health Tweets

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Freek Post
born in Vlaardingen, the Netherlands



Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

© 2014 Freek Post.

# Supporting users with credibility assessments of health Tweets

Author: Freek Post
Student id: 1229540
Email: freek@freekpost.com

**Abstract**

Twitter is one of the most popular social media platforms nowadays with more than 500 million tweets being send on a daily basis. A great variety of topics is discussed on Twitter and health has become a widely mentioned topic on Twitter. Users share updates regarding their own personal health situation and they use Twitter to look up health information. Inexperienced users have difficulty assessing the credibility of information contained in tweets. Additional visual and textual cues can be added to the tweets that enhance a users ability to determine the credibility. We define different types of feature sets to represent each tweet, which are then extracted and these features are used for automatic methods that can determine the credibility of health tweets. We narrow our domain to tweets about cancer and for each tweet we assess whether it's credible, neutral or non-credible. We also investigate the different types of cancer tweets and how the credibility differs between the categories of tweets. We evaluate our research by using a labeled ground truth obtained via the crowd-sourcing platform CrowdFlower. Our results show that automatically supporting users with credibility assessments of health tweets is feasible and can be employed in practice but still leaves room for improvement.

Thesis Committee:

| | |
|---|---|
| Chair: | Prof. dr. ir. Geert-Jan Houben, Faculty EEMCS (WIS), TUDelft |
| University supervisor: | Dr. ir. Qi Gao, Faculty EEMCS (WIS), TUDelft |
| Committee Member: | Dr. ir. Claudia Hauff, Faculty EEMCS, TUDelft |
| Committee Member: | Dr. ir. Felienne Hermans, Faculty EEMCS, TUDelft |

# Preface

First, I would like to thank Qi for his continued support and supervision during my project. His suggestions helped me greatly with my research. Our meetings together were inspiring and very helpful. Thank you to Alessandro and Claudia for providing useful feedback and criticism during our Omicron meetings and Claudia for supervising me in the last part of the research project. I would also like to thank Geert-Jan Houben for providing me with the opportunity to do my thesis at the WIS group, his suggestions for a research subject, and the feedback during the thesis project. Next I would like to thank Felienne Hermans for taking the time to be part of my thesis committee and advising me on my choice of beverages. Thank you also to all my friends who insisted that I finish this research project. Finally I would like to thank my parents and my sister for their continued support and patience during my thesis work.

Freek Post
Delft, the Netherlands
August 20, 2014

# Contents

**Preface** iii

**Contents** v

**List of Figures** vii

**1 Introduction** 1
- 1.1 Background . . . 1
- 1.2 Motivation . . . 1
- 1.3 Research objectives/questions . . . 2
- 1.4 Outline . . . 3

**2 Related work** 5
- 2.1 Health . . . 5
- 2.2 Credibility . . . 6

**3 Investigating the credibility problem** 9
- 3.1 Definition of credibility . . . 9
- 3.2 Data collecting and preparation . . . 10
- 3.3 Initial analysis . . . 11
  - 3.3.1 Data set division . . . 11
  - 3.3.2 Categories . . . 12
  - 3.3.3 Credibility . . . 13
- 3.4 Crowdsourcing with CrowdFlower . . . 14
- 3.5 Results . . . 15
  - 3.5.1 Credibility . . . 15
  - 3.5.2 Category . . . 16
  - 3.5.3 Limitations . . . 17

**4 Researching credibility of tweets** 19
- 4.1 Data collecting and preparation . . . 19
- 4.2 Ground truth . . . 20
  - 4.2.1 Results . . . 23
  - 4.2.2 Evaluation . . . 24

4.3 Feature selection . . . . . 24
4.4 Research methods and tools . . . . . 26
4.4.1 Classifiers/learning schemes . . . . . 26
4.4.2 WEKA . . . . . 27

**5 Credibility evaluation** **29**
5.1 Determining category . . . . . 29
5.2 Feature analysis category . . . . . 32
5.3 Determining credibility . . . . . 33
5.4 Feature analysis credibility . . . . . 36
5.5 Data analysis . . . . . 37
5.5.1 Relation between category and credibility . . . . . 37
5.5.2 Evaluating comments . . . . . 37

**6 Conclusions and Future Work** **41**
6.1 Conclusions . . . . . 41
6.2 Future work . . . . . 42

**Bibliography** **43**

**A Keywords** **47**

# List of Figures


3.1 Twitter Streaming API . . . . . 10
3.2 Example of a credible tweet . . . . . 14
3.3 Example of a non credible tweet . . . . . 14
3.4 Interface of the CrowdFlower credibility job . . . . . 15
3.5 Interface of the CrowdFlower category job . . . . . 16

4.1 Division of the data set . . . . . 20
4.2 Design of the CrowdFlower Category job . . . . . 21
4.3 Design of the CrowdFlower Credibility job . . . . . 22
4.4 CrowdFlower process flow . . . . . 22
4.5 Graphical User Interface of WEKA . . . . . 27

# Chapter 1

# Introduction

## 1.1 Background

The Web is a standard part of life for a large part of the population. We are continuously connected to the Web via our phones, tablets, computers and other electrical devices. People spend a lot of their 'online' time on social media websites, such as Facebook, Twitter and LinkedIn. Social media allows people to stay in touch with their friends and family, keep up to date on topics they are interested in, as well as consume and share content in a quick and simple way.

Twitter is an increasingly popular micro blogging social media service that allows sharing of all kinds of information via status updates. Users can follow each other and read the status updates which other users have shared. The simplicity of the Twitter service and the 140 character limit for status updates make Twitter easy and accessible for a wide audience. A wide variety of topics is discussed on Twitter ranging from news subjects to personal details about peoples lives. One such popular topic on Twitter is the topic of health[35][22]. An increasing number of users look up information about health related issues on the Web. People don't just read but also share personal information regarding their health online and Twitter is no exception[29]. Furthermore health organizations and doctors are using social media like Twitter to communicate and educate people about a wide variety of health topics[19].

## 1.2 Motivation

People look up information about health related topics on social media such as news about diseases, advice on what do, research regarding new treatments and to read about experiences of other people. The fact that people share their experiences allows other users of Twitter to use the platform to read and learn from each other. For example, sufferers from insomnia can use Twitter to find messages from people with insomnia who post tips on how they deal with the infliction[22]. It is a free and open, ever increasing catalog of information to which any person can contribute. Inexperienced users of the Web however might have difficulty with assessing the credibility of information compared to more experienced users[26]. When we talk about credibility we use the definition "'offering reasonable grounds for being believed"'[4] just like

Castillo et al.[11] as we discuss in Section 3.1. We therefore make a distinction between a tweet that is credible and a tweet that is true. In this research we focus on the credibility of a tweet and not whether the content of a tweet is true or not. Research has been done to determine the credibility of tweets[11][23] but not specifically for the topic of health. We believe that we can assist the users of Twitter in this effort by automatically determining the credibility of tweets in the health domain. In order to narrow the scope of the thesis we have decided to restrict ourselves to a more narrow domain for which we choose cancer. Cancer affects a large number of people every year with over 1.6 million expected new cancer cases in the US alone[1] and it is a widely discussed topic on Twitter which makes it a good research subject. If we can support users with credibility assessments for the topic of cancer we can then apply this research to the general health domain.

## 1.3 Research objectives/questions

Our main research goal that we aim to achieve with this research project is:
*Support users with making credibility assessments for tweets in the health domain specifically cancer.* In order to achieve this goal we have distilled four research questions that we need to answer.

**1. How do we find relevant tweets about a disease (cancer)?**

First we need to get a set of relevant tweets about cancer for which we can then assess the credibility of these tweets in order to support the user. When we say relevant we mean tweets that mention the disease cancer in any way and not tweets that mention other meanings of the word cancer.

**2. How can we automatically determine the type or category of a tweet?**

There are different types of tweets posted about cancer on Twitter ranging from personal experiences to research about a disease. We want to investigate whether the type of a tweet has an effect on the credibility perception of users. We can then use the type of the tweet in the process of assessing the tweet credibility. We will therefore divide the tweets into different categories such as advice, news, research and personal tweets.

**3. What features are important for determining credibility of tweets in the health domain?**

To gain a better understanding on what people find important to determine the credibility of a tweet we will investigate which features have the biggest impact on this process.

**4. How do we automatically determine the credibility of tweets in the health domain?**

Finally we will research how we can automatically determine the credibility of tweets about cancer and use this to assists users in making their own credibility assessments.

## 1.4 Outline

The rest of this thesis will be structured as follows. In Chapter 2 we give an overview of the related work that has been done in this field. Next in Chapter 3 we do an initial analysis on Twitter to investigate whether our perceived credibility problem exists. In Chapter 4 we collect a data set of tweets related to cancer to investigate, from which we create a ground truth through the use of the crowdsourcing platform CrowdFlower. Furthermore we describe the features that will be used and we discuss the available methods to determine the credibility. Chapter 5 presents the results of the evaluation of automatically determining the credibility of health tweets and we discuss our findings. Finally in Chapter 6 we conclude the thesis with our conclusions and recommendations for future work.

# Chapter 2

# Related work

Given the widespread use of social media by large numbers of users nowadays and the impact social media has had on the social interactions between people, the topic of social media is an interesting field of research. Social media and social networks are used for a widespread array of applications and contain large amounts of data which researchers can use for analysis and experiments. One of the more popular social media is Twitter. Twitter is a microblogging service that allows people to send status updates of up to 140 characters which are called tweets. Users of Twitter can follow other Twitter users to receive the tweets they post. Users can also search Twitter for specific topics of interests ranging from personal statuses, news, events and even health. Research has been done to detect the different types of information on Twitter, such as filtering out uninteresting content[8], detecting earthquake tweets[32], finding incident information[7] and filtering users timelines with topics of their interest[16].

## 2.1 Health

People talk about their own health issues and illnesses, the symptoms they have, how they feel about it and what they do to treat their diseases[29]. Since Twitter is used by hundreds of millions of users every month with 500 million tweets being send out every day[5], there's a lot of health information that can be investigated. Paul et al.[29] analyzed over one and a half million tweets containing user generated health information. In their research they found over a dozen ailments such as insomnia, obesity and allergies as well as symptoms people have and the treatments they use to combat these ailments. They improved the Ailment Topic Aspect model(ATAM)[30] which creates disease information from tweets and they applied ATAM for analyzing symptoms and medications, following illnesses over time and localizing them by region, and for tracking behavioral risk factors. Powel et al.[22] focus on an individual illness for which they chose the sleeping disorder insomnia. A content analysis was performed on a set of tweets containing the words insomnia which revealed two categories of tweets. People described how they are coping with insomnia and what their experience with insomnia is. Based on these results a proposal is made to improve access to social support as well as therapy for sufferers from insomnia by creating a social network platform.

Social media are not just used by the patients but health organisations and doctors are beginning to embrace the social media platforms as new ways to communicate and interact with patients[19]. One example of this is Hello Health, a small medical practice that offers a secure private social network through which patients can ask questions and communicate with their doctor directly for a monthly fee. The field of e-health comes with both benefits as well as downsides and risks. These benefits include many to many instead of one to one communication, patient empowerment, easier communication and improved patient happiness. However there are also downsides and risks with e-health such as privacy issues, maintaining standards of care as well as costs. Research has been done to investigate what the impact is on health outcomes from health messages in social networks and methods have been discussed to understand why some health messages are trusted more by users[27]. Another study looked at health messages on Twitter to use as an early detection mechanism for the outbreak of diseases and to provide public health officials with useful information. Recommender techniques are applied to rank health tweets based on the interests of the public health officials with regards to disease, symptoms and location[14].

## 2.2 Credibility

Studies have been performed to assist users in assessing the information on social media[13] by providing visual cues. Evaluations show that users found these visual cues helpful and allowed them to more easily assess information found on social media. Castillo et al.[11] focus on developing automatic methods to determine the credibility[15] of groups of news tweets obtained from "'trending topics"'. They define four types of feature sets (message-based, user-based, topic-based and propagation-based) which are used to describe each topic set of news tweets. Classifiers that are trained using the feature sets achieve a 86% accuracy distinguishing between credible news topics and non-credible ones using a labeled ground truth data set. Another study presents three models for determining the credibility of topic specific tweets[23]. The social model looks at credibility on the user level, the content model which focuses on the tweet content and the hybrid model which combines these two models to make a credibility assessment. Results from the evaluation of these models show that the social model outperforms the other two models with an accuracy of 88%. Research has found that the best features to determine credibility from are URLs, mentions, retweets and the length of the tweet[28].

Work has also been done to find credible users that provide relevant information for a certain topic[10]. Credibility is viewed as a combination of trust and expertise, with expertise being based on a userś messages. Social network users are identified and ranked on relevance for a certain topic by combining text search and social network analysis. A similar system is Cognos by Ghosh et al.[17] through which topic experts on Twitter can be found. Cognos uses the List feature of Twitter to determine experts for topics from lists that other users have filled themselves. The crowdsourced approach of Cognos can compete with Twitters own "'Who To Follow"' functionality.

Morris et al.[26] investigate what factors users are influenced by to make credi-

bility assessments. They found that users rely on factors such as whether a tweet has been retweeted, the user image and whether a username is topically related for making credibility judgements. Improvements are suggested for search engine interfaces to include these features when displaying microblog updates outside of the social media platform. A similar study has been done that focuses on enhancing users ability to assess the credibility of Web pages[33]. Visualisations are added to Web pages that show popularity of the page among the general public and experts, any awards the site has won, as well as where and when people have accessed the Web page from. Evaluations show that users are able to make more accurate credibility assessments of Web pages by having these features. According to Kakol et al.[25] credibility is subject to bias which they attest to acquiescence bias which causes the users to be overly positive when rating pages. Rafalak et al.[31] looked at the psychological factors that influence credibility evaluations of web content. They looked at several psychological traits like trust, conformity, risk taking, need for cognition and intellect of the three types of raters: under-raters, average-raters and over-raters. Their results show that people with a high general trust level and a low risk taking level are more likely to overestimate the credibility of websites. They explain that a possible reason for this is that choosing credible to rate a website is seen as the safe option.

The work of Gupta et al.[18] focuses on distinguishing fake from real images during Hurricane Sandy to prevent the false information from causing chaos and panic among people. Their investigation shows that 86% of the tweets spreading fake images are retweets and 90% of the retweets came from 0.3% of the users. Thus a large part of the fake images came from a very small portion of the total users. A Decision Tree classifier is able to achieve a 97% accuracy in detecting fake images from real images by using tweet based features.

# Chapter 3

# Investigating the credibility problem

In order to find out whether there really is a problem with the credibility of health tweets and to gain a better understanding regarding this research project we start with an initial investigation. Existing research has shown that it is possible to distinguish newsworthy tweets from other messages and to automatically determine the credibility of newsworthy tweets[11]. Various models have been created and evaluated to assess the credibility of tweets[23]. We will use the existing research and apply the findings to the health domain.

First we define the notion of credibility. Next we create a sample data set by collecting tweets about cancer from Twitter. We will do an initial analysis on this sample data set manually to see if we can get any knowledge that will help us during the rest of the research project. Then we use the crowdsourcing platform CrowdFlower to find out what people think of the credibility of our set of tweets. Finally we discuss the results that our initial investigation has yielded.

## 3.1 Definition of credibility

In this research we focus on the credibility of a tweet and not whether the tweet is true or not. Existing research has been done in this field and various definitions of credibility have been given. In 1999 Fogg and Tseng[15] investigated the various aspects of computer credibility. They defined four different types of credibility as well as three strategies for credibility evaluation. Their work has been used by many as the basis for further research about credibility. They found that believability can be used synonymously with credibility and that credibility is a perceived quality which is based on trustworthiness and expertise. Next Kang et al.[23] and Castillo et al.[11] both have similar definitions for tweet credibility: an indication that the tweet contains believable information. Kang et al.[23] also define social credibility which they describe as the believability of the used based on their standing in the social network. The definition of credibility which we will use throughout this thesis is the one from Merriam-Webster[4] which defines it as: offering reasonable grounds for being believed.

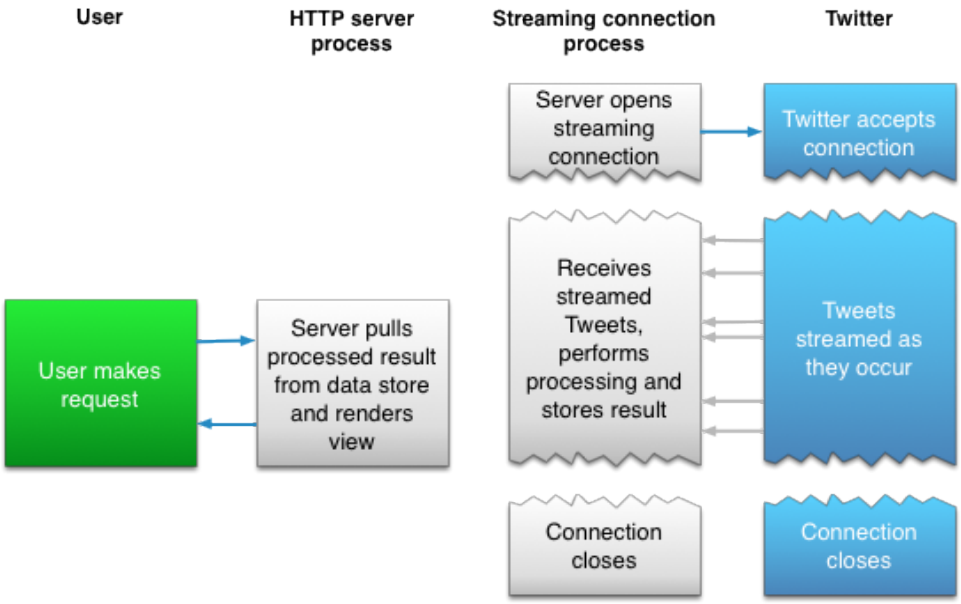



Figure 3.1: Twitter Streaming API

## 3.2 Data collecting and preparation

For our initial investigation we want to create a sample data set on which we can do an analysis to get a better idea about the research and the domain. We use the Twitter Streaming API to collect a set of tweets related to cancer. Twitter provides streaming API's (application programming interface) that differs from their REST API. With the streaming API an application maintains a persistent open HTTP connection with Twitter through which tweets are pushed as they are created as can be seen from Figure 3.1. The advantage of using the streaming API compared to the REST API is the fact that we don't have to continuously poll the REST API to get more tweets. Twitter offers 3 streaming API'S:

**Public streams**
: Streams of the public data that is available on Twitter which can be filtered with keywords and specific users.

**User streams**
: The streams with tweets for a single user on Twitter.

**Site streams**
: Similar to the User streams but for multiple users simultaneously.

We use the Public stream API since we want data from multiple users and we will filter by using keywords related to cancer. We create a list of 22 keywords that are related to cancer by manually selecting common cancer terms from the cancer.gov website[2]. The list consists of terms such as types of cancer, treatments and cancer specialists as can be seen in table 3.2. The terms of this list was chosen by the author

Table 3.1: Cancer terms used in collecting tweets from the Twitter streaming API

| | | |
|---|---|---|
| biopsy cancer | bladder cancer | breast cancer |
| chemotherapy | colon cancer | doctor cancer |
| endometrial cancer | Hodgkin lymphoma | kidney cancer |
| leukaemia | lung cancer | mammogram |
| mammography | mastectomy | melanoma |
| oncologist | pancreatic cancer | prostate cancer |
| rectal cancer | skin cancer | thyroid cancer |
| tumour cancer | | |

based on several searches on Twitter with the "‘cancer"’ keyword. With these initial searches we found various tweets which included personal ones from people with cancer as well as organisations and doctors giving advice. Each tweet that we collect will contain all keywords of one of the cells in the table. We chose not to monitor for tweets containing only the keyword âcancerâ as this returned a lot of tweets that were unrelated to the disease cancer. An explanation for this is the fact that cancer has multiple meanings in the English language such as the astrological sign Cancer[3].

We collected a set of 81434 tweets from the Twitter Streaming API that contained one or more of the cancer keywords over a two week period from 27th of August 2012 till 17th of September 2012. Twitter returns each tweet in JSON (JavaScript Object Notation) format which contains information about the tweet itself including entities the tweet contains such as hashtags, URLs and media, as well information about the user of the tweet. Each tweet and user is parsed and stored in a database as well as the raw JSON data of each tweet.

## 3.3 Initial analysis

### 3.3.1 Data set division

Now that we have collected our sample data set we will perform an initial analysis on it. We want to test our hypothesis that there are health tweets which people view as not credible. We start by looking at the division of the set of tweets. The data set consists of:

**Tweets**
The standard form of a Twitter status update

**Retweets**
A tweet in which a user copies the content of another tweet with RT @username in front of it to indicate he or she agrees with the tweet or to give the message a greater reach. Twitter has now added an official "‘Retweet"’ button which shows the original tweet "‘as is"’ on the retweeting users timeline. However there are still users who use the unofficial method of retweeting.

Table 3.2: Top 10 most used hashtags in our collected sample data set

| Hashtag | Amount |
|---|---|
| TWmeetDayna | 1826 |
| cancer | 1435 |
| marijuana | 990 |
| YouWereHotUntil | 971 |
| WWENOC | 633 |
| projectpink | 625 |
| breastcancer | 616 |
| RiseAboveCancer | 576 |
| Health | 416 |
| leukemia | 416 |

**Replies**

A reply is a tweet that's a response to another tweet, similar to a reply of an email message.

Each tweet can contain zero or more hashtags and/or URLs. If we look at the top 10 hashtags of our sample data set we can see that the most used hashtag is #TWmeet-Dayna which is used for a Twitter effort to help a student with leukemia to meet her favorite band. The number 2 hashtag is #cancer which we would expect given the fact that we are querying tweets with keywords related to cancer. We can also see that most of the users are only linked to a single tweet, with 11 users having placed more than a 100 tweets in the time that we monitored. Upon further inspection of these users we see that 6 of them are doctors or individuals tweeting about cancer news and advice, while the other 5 users are spamming the same sort of tweets repeatedly.

### 3.3.2 Categories

Browsing through the data set we find that there are different types of tweets about cancer. We want to investigate how the credibility differs between the different types of tweets as this could then be used in determining the credibility. In order to do that we divide the tweets up into different categories. We create a list of categories by going through a subset of the data set and we create a new category based on the type of the tweet if the tweet does not fit an existing category according to the authors analysis. Initially we came up with the following list of 9 categories.

1. Research
2. News
3. Events
4. Advice
5. Chat
6. Personal experiences

7. Questions
8. Sale/Fundraise
9. Non-relevant

However we decided that our initial list was too broad since we found that quite often a tweet could belong to multiple categories. Upon further inspection we find that we can narrow the list by merging similar categories together which resulted in the following list of 5 categories.

1. **News**
    a) News
    b) Research
2. **Advice**
    a) Advice
3. **Event/Fundraise**
    a) Event
    b) Sale/Fundraise
4. **Personal**
    a) Personal
    b) Questions
5. **Other**
    a) Chat
    b) Non-relevant

### 3.3.3 Credibility

Finally we look at the credibility of the collected tweets to find out if there are tweets related to cancer which can be classified as not credible. As stated in Section 3.1 we use the definition "'reason enough to be believed"' when we talk about credibility. Just as we did with the categories in the previous section we browse through the data set and manually inspect the tweets. During this process we noticed that there are tweets for which we believe they can be viewed as not credible. To demonstrate our findings we give an example of a credible tweet in Figure 3.2 and an example of a non credible tweet in Figure 3.3. We believe the tweet in Figure 3.3 is not credible based on the authorś personal experience that if a person has cancer they are usually treated with chemotherapy and never with lime juice. In order to confirm our initial findings we will use the crowdsourcing platform CrowdFlower in the next section to ask other people how they would rate a sample subset of tweets.




Figure 3.2: Example of a credible tweet




Figure 3.3: Example of a non credible tweet

## 3.4 Crowdsourcing with CrowdFlower

CrowdFlower is a crowdsourcing platform that allows users to submit tasks which other people can do for a small monetary compensation. These tasks called Human Intelligence Tasks (HIT) are tasks that a computer is not well suited to perform but are usually easy for human beings. Our choice to use CrowdFlower in our initial study has three reasons. First we want to investigate how people would classify our sample set of tweets with regards to credibility and category. The goal is to find out if other people perceive some tweets as non credible and to determine whether they label the tweets with the same category as the author. Second we want to determine whether CrowdFlower is suited for annotating a larger data set for the main research project. Finally by doing small initial tests with CrowdFlower we can improve our design of the tasks which will improve the results we obtain for the main research project.

We set up our first job on CrowdFlower to ask people to judge the credibility of a set of tweets. We used a set of 20 tweets and asked for 5 judgements per tweet which means each tweet will be judged by 5 different persons as recommended by Ipeirotis et al.[21] for better quality results[36]. For each tweet a person can choose between four levels of credibility: (a) almost certainly true, (b) likely to be false, (c) almost certainly false, and (d) i can't decide. These options for the credibility level were based on the work of Castillo et al[11]. We used a design for each tweet similar to that of Twitter to mimic the actual interface as close as possible. Figure 3.4 shows a screenshot of the job interface the users were presented with.

Next we used CrowdFlower to ask users to judge the category of a set of tweets. We create a separate job for the category judgements instead of combining it with the credibility job as this might complicate things for the people doing the tasks and thus influence the results. Based on the process of our credibility job, we made two improvements for the category job. First we make use of CrowdFlowers "'Gold data"'

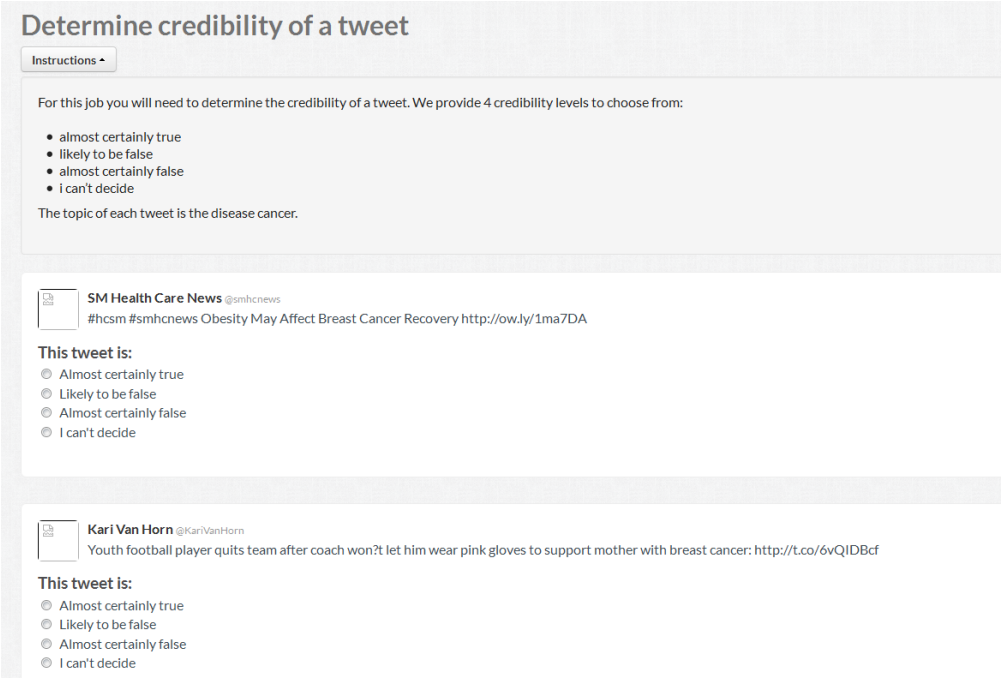



Figure 3.4: Interface of the CrowdFlower credibility job

feature which allows you to add test data to your data set for which the answer is already given. The gold data is used to prevent people who do these tasks from just filling in random answers since we use multiple choice questions[34]. Next we also asks the users to explain their decision in the form of a comment to gain a better understanding of their choices. A screenshot of the job interface can be seen in Figure 3.5.

## 3.5 Results

### 3.5.1 Credibility

We now look at the results of our CrowdFlower experiment for both the credibility task and the category task. With the credibility task we want to determine whether there is a problem with the credibility of health tweets. Of our small set of 20 tweets, 3 of the tweets are labeled as "'Likely to be false"' or "'Almost certainly false"' by the CrowdFlower workers which are shown in Table 3.3 along with the confidence score. The confidence score describes the agreement between workers about the result for each tweet. For each possible answer, which in our case is the credibility level, a confidence score is calculated based on the trust scores of the workers who picked the answer. The answer with the highest confidence score is then taken as the result of the task. The higher the confidence score the more agreement there is for the final answer. The first result in Table 3.3 has a confidence score of 0.63 which means that there is reasonable agreement about the answer. Of the 5 contributors for labeling this tweet, 3 of them chose "'Almost certainly false"', 1 chose "'Likely to be false"' and

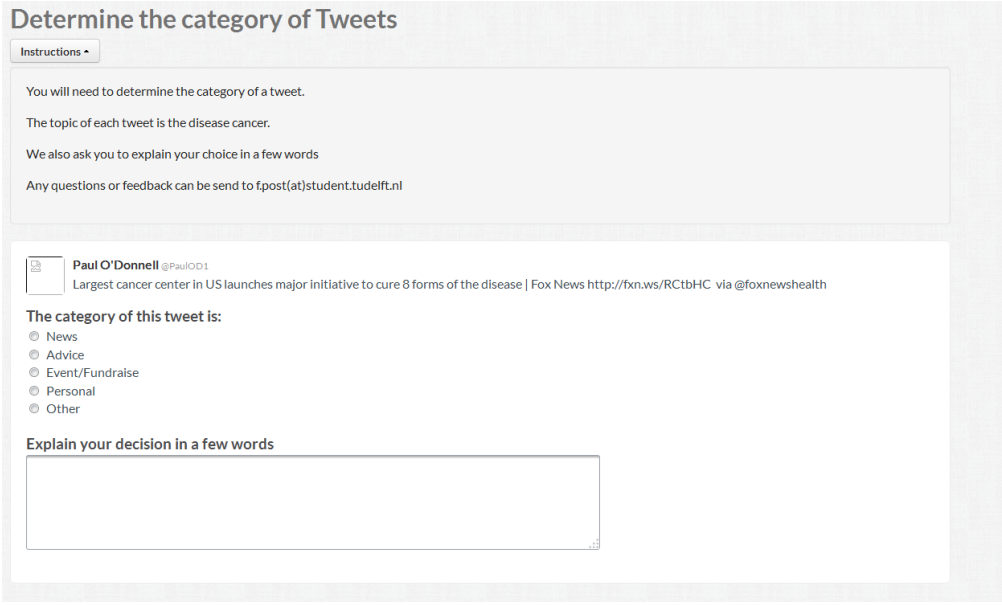



Figure 3.5: Interface of the CrowdFlower category job

| Credibility level | Confidence score | Tweet text |
|---|---|---|
| Almost certainly false | 0.6309 | Chemotherapy Makes Cancer Worse? http://t.co/2zmCaiee |
| Likely to be false | 0.5741 | Lung cancer kills more women than breast cancer, so why the funding gap?: ATLANTA ?ÃÂ¿The news hit Amy Waggoner hard. http://t.co/nRY5hsI2 |
| Likely to be false | 0.3943 | Orgasms cure migraines in women and prostate cancer in men.. |

Table 3.3: Tweets labeled as Non-credible on CrowdFlower

1 chose "'I can't decide"'. For the second result there is also reasonable agreement about the answer with a confidence score of 0.57. The third result has less agreement about the answer "'Likely to be false"' with a lower confidence score of 0.39. Of the 5 contributors, 2 chose "'Likely to be false"', 2 chose "'Almost certainly false"' and 1 chose "'I can't decide"'. We see that there is ambiguity between the "'Almost certainly false"' and "'Likely to be false"' options and we will investigate this further in our main experiment. The initial research has shown that there are health tweets which people perceive as Non-credible. Based on these results of the initial investigation we decide to continue with our research project.

### 3.5.2 Category

Next we look at the result of our category task for which we want to see how people label the tweets with a category. The small set of 20 tweets was labeled by the workers on CrowdFlower as well as ourselves so we can compare the results. The tweets in the

| | | |
|---|---|---|
| Advice | 0.3741 | You may have to do things for everyone else today, which makes... More for Cancer http://bit.ly/yk3b9m |
| Personal | 0.581 | Cancer people might be some of the most dramatic individuals you will ever encounter. |
| Event/Fundraise | 0.5929 | I've just made a donation to Stand Up To Cancer - you should too! http://shar.es/5uthD |

Table 3.4: Tweets with low confidence categories

categories of News, Event/Fundraise and Personal are labeled similar by the CrowdFlower workers as by us. The tweets which we labeled as Advice and Other were not all labeled similarly by the CrowdFlower workers. If we look at the individual labeling and the confidence scores of the tweets in these categories we see that there is less agreement among the workers as well. Several of these tweets with low confidence scores are shown in Table 3.4.

### 3.5.3 Limitations

Our tasks that we set up on CrowdFlower for labeling our small set of tweets with a credibility level and a category is quite simple. We do not take extensive anti-spam measures or validation checks since we simply want to determine if there is a perceived credibility problem and we do not mind if there are a few spam results in between the total result set. We still ask for multiple judgements per tweet to ensure a minimum of validity for our results. We collected our data with a small set of cancer related keywords since our goal here was to obtain a sample data set with tweets about cancer. We do not yet need a diverse data set with all types of cancer for this initial investigation.

# Chapter 4

# Researching credibility of tweets

In the previous Chapter we performed an initial investigation to determine whether there is a problem with the credibility of health tweets. We used CrowdFlower to ask people to label our tweets with a credibility level as well as a category. From this investigation we found that there are tweets which people rate as not credible and we therefore continue with our research. In order to automatically determine the category and credibility of a set of tweets we will first create a data set including a ground truth as a baseline. Next we determine the features of the tweets that will be used in the process of classifying the tweets. Furthermore we will discuss the research methods and tools that we will use in our research project.

## 4.1 Data collecting and preparation

For our main research project we create a new data set of tweets related to cancer. As in Chapter 3 with our initial investigation we will use the Twitter Streaming API to collect the tweets. We manually compile a list of keywords from the Cancer.gov website[2] where we include all the list of all cancer types. Our list of cancer keywords can be divided into four categories: (i) cancer types, (ii) treatments, (iii) professions, and (iv) related words. We also include the keyword "'cancer"' since we want to include non related tweets in our data set to investigate whether we can correctly determine the category of our data set of tweets. Our final list consists of 398 keywords related to cancer which can be found in Appendix A. This final list is a lot larger than the list from our initial investigation in Section 3.2 since we now want to get a diverse data set of cancer tweets to research. The list in Section 3.2 was only used to collect a small data set of cancer tweets which did not need to include a varied data set for our initial investigation.

Over a period of 1 month in the month of September 2012 we collected a set of 5325295 tweets via the Twitter Streaming API. The tweets obtained from Twitter are represented in JSON format which we convert to separate entities for the tweets, users, hashtags, URLs, places and media. The collected data set consists of regular tweets, retweets and replies with a division as shown in Figure 4.1. From this data set we want to create a subset that we can get people on CrowdFlower to rate in order to create a smaller data set of 1000 tweets we can use as a ground truth for both category and credibility. We query the database for a random set of tweets and filter these on

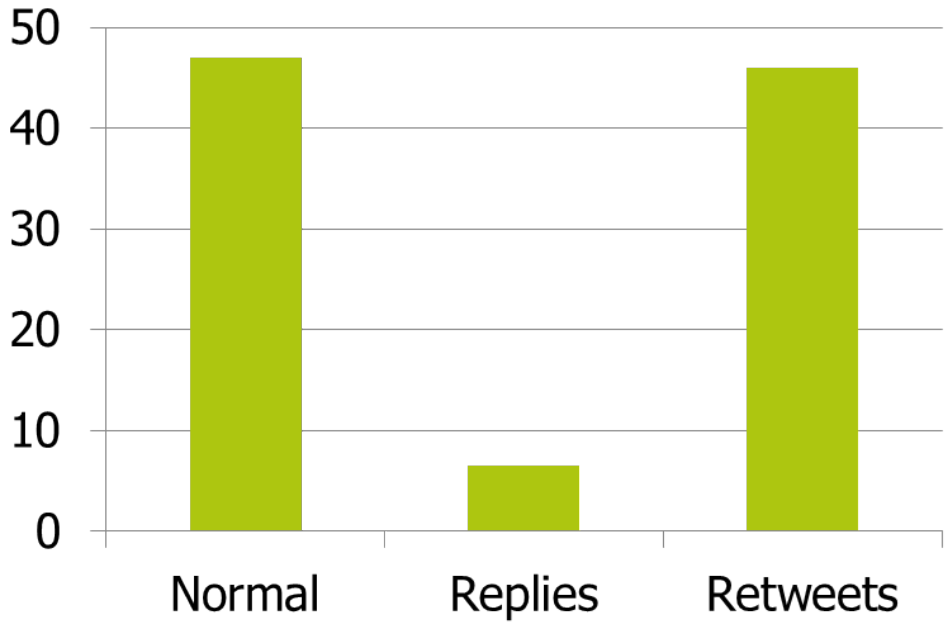



Figure 4.1: Division of the data set

keywords. We exclude all the retweets because these retweets are basically copies of other tweets which are also in our collection of tweets and these retweets make it harder to select a varied subset of tweets. Initially we selected a random set of 1000 tweets several times. However upon inspection by the author we ended up with subsets that had tweets which all seemed credible or were all of the same category. We want a subset that is varied with regards to credibility and category in order to use it as a ground truth. To accomplish this we use keywords that are common for each category to query for tweets which we then add to our smaller data set and we manually balance this data set to ensure there are enough non credible tweets in this data set.

## 4.2 Ground truth

We now have a smaller data set of 1000 tweets which we will use to create a ground truth data set. The ground truth data set will be used to evaluate our automatic methods for determining the credibility and category of health tweets. To create the ground truth we will use CrowdFlower again since our initial research showed it suitable for labeling a data set. CrowdFlower allows for multiple people to label our collection of tweets for a small monetary compensation which saves us time and effort.

Our design of the jobs on CrowdFlower is similar to those in the initial research but we have made several improvements from the results of that initial research and existing research[26]. As mentioned we want to mimic the Twitter design as close as we can and we therefore now show the amount of retweets as well as the following/followers amount for each tweet. These two features of the tweet can be important as people who view the tweet might view tweets which have been retweeted by several

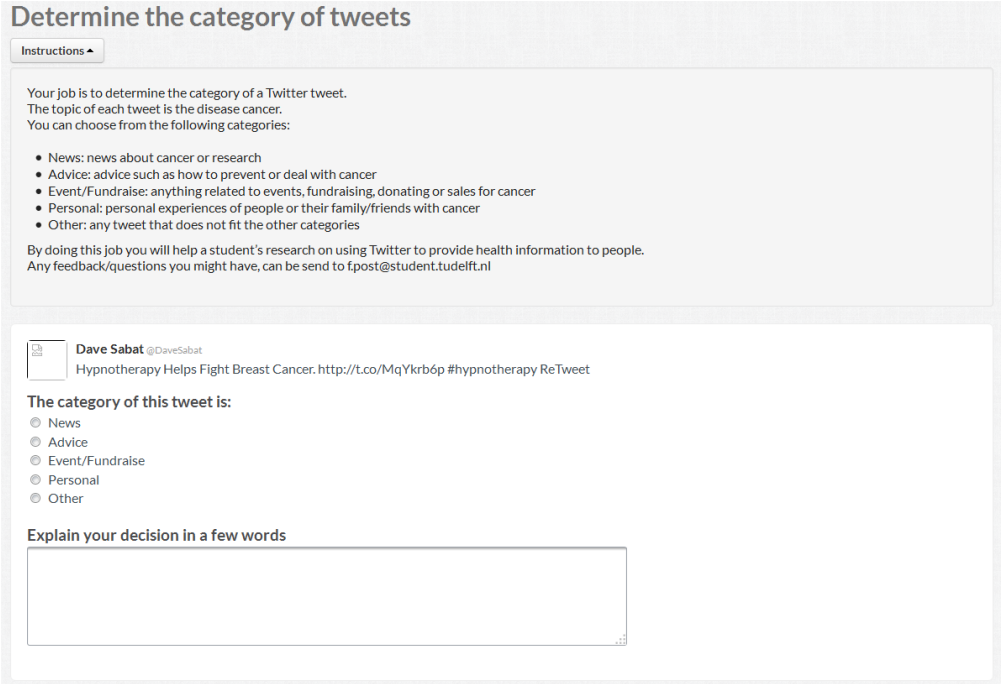



Figure 4.2: Design of the CrowdFlower Category job

people as more credible[23]. Next we also ask people to explain their decision for choosing a particular category or credibility level such as the content of the tweet or the amount of retweets. We ask for 5 judgements for each tweet as opposed to the the 7 judgements used by Castillo et al[11]. However Castillo et al.[11] rate a set of 10 tweets whereas we rate each single tweet and we find 5 judgements is the right amount for our process as 5 judgements is also recommended by Ipeirotis et al.[21].

We create separate jobs for both category and credibility to keep the jobs as simple as possible so that we increase our chance of people performing the work for us and to reduce the chance of influencing the results by combining the jobs. Gold data is used again to prevent people from randomly filling in answers without actually reading the questions. The gold data in combination with 5 judgements per tweet provides a higher chance of getting better quality data[21][34][36]. We first set up a job for category with our set of 1000 tweets and 81 tweets of gold data with the final design for the category job shown in Figure 4.2. Since the category job on CrowdFlower took a long time to complete we decided to split up the job for credibility into 6 smaller jobs of each 200 tweets with 10 tweets per job as gold data. The design for the credibility jobs is similar to the category job on CrowdFlower as shown in Figure 4.3. In our initial test with CrowdFlower in the previous chapter we saw that workers found it difficult to choose between the options "'Almost certainly false"' and "'Likely to be false"'. Our decision is to replace these two options with the options Non-credible and Neutral to provide a more clear distinction among the available options similar to what Castillo et al.[11] did.

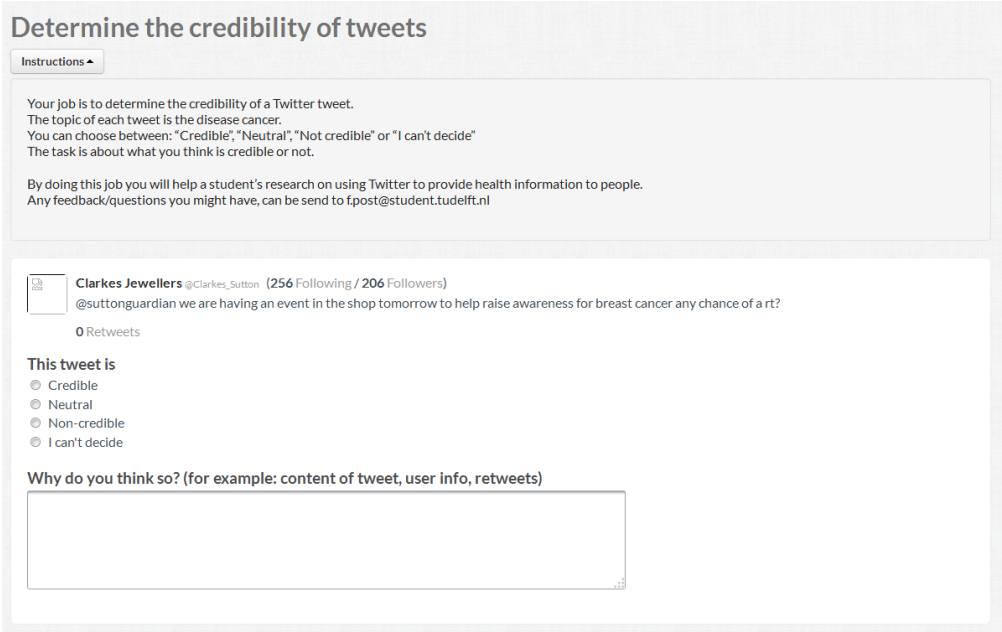



Figure 4.3: Design of the CrowdFlower Credibility job

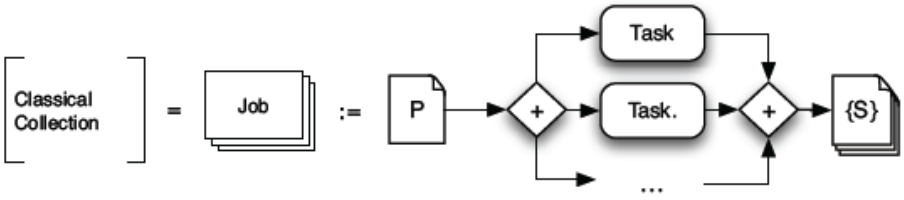



Figure 4.4: CrowdFlower process flow

We want to support the users with determining the credibility of tweets and as such our target audience are regular people of all types who use the internet and who may vary in skill from the very technical users to those very little technical knowledge. We therefore want to use what the average people think what is credible or not and use that in creating our ground truth. The framing of our question and the options that the people have are simple so a regular Web user can answer them.

The flow of our process on CrowdFlower is as follows. We create a job on CrowdFlower by uploading our data, uploading our gold data and setting the parameters of the job such as number of judgements, monetary compensation and crowdsourcing channels. When this is complete CrowdFlower posts the job on the chosen crowdsourcing channels which in our case is Amazon Mechanical Turk(AMT). Workers on AMT find our job, perform the task and if they pass the gold data questions their answer is accepted. Once the job has the required number of judgements for each tweet the job is complete. The entire flow of this process is shown in Figure 4.4

Table 4.1: Division of the credibility ground truth

| Credibility level | Amount | Percentage |
|---|---|---|
| Credible | 722 | 60% |
| Neutral | 276 | 23% |
| Not credible | 195 | 16% |
| Can't decide | 7 | 1% |

Table 4.2: Division of the category ground truth

| Category | Amount | Percentage |
|---|---|---|
| Advice | 112 | 11% |
| Event/Fundraise | 191 | 19% |
| News | 362 | 36% |
| Personal | 275 | 28% |
| Other | 59 | 6% |

### 4.2.1 Results

#### Credibility

We now look at the results of our credibility task on CrowdFlower. In total we have labeled 1200 tweets with a credibility level via the workers on CrowdFlower. The division of these tweets with regards to the credibility level is shown in Table 4.1. As we discussed in Section 3.5 the confidence level describes the agreement between the contributors for the final answer. The higher the confidence level the more agreement there is between the workers and the more confidence CrowdFlower has in the answer provided. There are 280 tweets with a confidence score below 0.5 which indicates a below average confidence for the tweet credibility level. Of these 280 tweets, 114 tweets are labeled Credible, 57 Non-credible and 109 Neutral. We can see that, with regards to the total amount for each credibility level, there is more disagreement for the Neutral class. If we look at the individual judgements for these tweets we can see that judgements for the Neutral class vary more between all three classes since it is the middle option.

#### Category

Next we look at the results of the category task on CrowdFlower. In total we have labeled 1000 tweets with a category via the workers on CrowdFlower similar to the credibility task. The division of these tweets with regards to the category is shown in Table 4.2. There are 104 tweets with a confidence score below 0.5 which indicates a below average confidence for the tweet category. Of these 104 tweets there is about an equal division of all categories in this subset showing that overall the Advice and Other category have the most disagreement with regards to the total amount for each category.

### 4.2.2 Evaluation

We now evaluate our experience with the use of the CrowdFlower platform to label data for our research. CrowdFlower can be an effective way to label your data but it requires setting up the job properly to ensure the accuracy of that data. Several mechanisms are provided by CrowdFlower as a form of quality control such as multiple judgements per unit[21][36], test data[34] and asking for feedback[9]. Our experience has been similar as those found in other researches. As mentioned jobs can also take a long time to complete depending on the number of units, judgements, monetary compensation and difficulty of the work. Splitting a single job up into several smaller jobs can reduce the time it takes for the work to be completed. Activity on our task was highest when just posted and slowed down over time. To combat that issue we republished our job via CrowdFlower when task completion slowed down as advised by Ipeirotis et al.[20]. In our category job we made the mistake of changing a label after having set up our gold data which caused us to have a test question which did not have a correct answer. The result of this was that several of our workers were angry that their work was rejected even though they answered the test questions correctly. CrowdFlower does not prevent these mistakes and care must be taken to avoid them. Finally we only used workers from the US, UK, Belgium and The Netherlands since Kazai et al.[24] found that European and US workers provide better quality labels than Asian workers.

## 4.3 Feature selection

The purpose of this research is to determine the level of credibility and category of a tweet. We do this with a list of features that we obtain from our data set. We have 3 types of feature sets which we use for determining the credibility and category: tweet features, user features and network features. These lists of features are mainly taken from Castillo et al.[11] and Kang et al.[23]. We do not use the topic features as described in Castillo et al.[11] since we only look at individual tweets instead of a set of tweets.

**Tweet features** are features of the tweet itself, for example the number of characters of a tweet, or whether the tweet contains a question or an exclamation mark. Beside the textual features we also use information such as whether the tweet contains a media item like a photo or video, whether a tweet has a location and when the tweet is placed

**User features** are features of the user who created the tweet, such as the number of followers of that user, the amount of statuses that user has created and whether the user is verified. We also use information the user has provided himself such as the description in his or her profile.

**Network features** are a small set of features of the Twitter network such as the amount of retweets and whether a user has set the tweet as a favorite.

Table 4.3: Tweet features

| Feature | Description |
|---|---|
| Number of characters | Length of tweet in characters |
| Number of words | Length of tweet in words |
| Number of urls | Number of urls |
| Has question mark | Has question mark |
| Has exclamation mark | Has exclamation mark |
| Has multiple questions/exclamation marks | Has multiple questions/exclamation marks |
| Number of uppercase chars | Number of uppercase chars |
| Number of Mentions | Number of Mentions |
| Number of Hashtags | Number of Hashtags |
| Has positive emoticon | Has positive emoticon |
| Has negative emoticon | Has negative emoticon |
| Is Weekday | Is Weekday |
| Has pronoun first\|second\|third | Has pronoun first\|second\|third |
| Number of positive words | Number of positive words |
| Number of negative words | Number of negative words |
| Sentiment score | Sentiment score |
| Number of intensifiers | Number of intensifiers |
| Number of swearwords | Number of swearwords |
| Is only URLs | Is only URLs |
| Has media | Has media |
| Has location | Has location |
| Is Daytime | Is Daytime |
| Is reply | Is reply |

Table 4.4: User features

| Feature | Description |
|---|---|
| Registration age | Registration age |
| Statuses count | Statuses count |
| Followers count | Followers count |
| Friends/Following count | Friends/Following count |
| Is verified | Is verified |
| Has description | Has description |
| Has url | Has url |
| Has default profile image | Has default profile image |
| Has default profile | Has default profile |
| Has location | Has location |
| Has topical username | Has topical username |
| Type of user image | Type of user image |
| User tweets frequently | User tweets frequently |
| User bio contains topic information | User bio contains topic information |
| Followers/friends ratio | Followers/friends ratio |
| Favourites count | Favourites count |
| Listed count | Listed count |

Table 4.5: Network features

| Feature | Description |
|---|---|
| Is Favorited | Is Favorited |
| Number of retweets | Number of retweets |
| Is Retweet | Is Retweet |

## 4.4 Research methods and tools

In the previous section we have defined a set of features which we can use to determine the category and credibility of tweets about cancer. We now describe the research methods and tools that we employ to create a model for deciding the category and credibility of each tweet. We use a supervised learning approach and we train our classifiers on the features defined in section 4.3 using the ground truth data set of section 4.2. Our approach is similar for both credibility and category but the optimal learning scheme may differ between these two.

### 4.4.1 Classifiers/learning schemes

We train a supervised classifier using the ground truth data set created from CrowdFlower as a training set. We train separate classifiers for credibility and category. We test different types of classifiers and compare the results of these to each other to find the one best suited for our tasks. We test the J48 tree-based learning scheme so we can compare the results to those of Castillo et al.[11] and Kang et al.[23] as well as other classifiers such as the Logit boost and the Logistic classifier.

As mentioned in section 4.2 we have four labels for our data: Credible, Neutral, Non-Credible and Can't Decide. We try out different strategies with regards to these labels and compare the performances of each strategy. We will specifically focus on the effect of the Neutral class and the effect on the performance it has when we leave the class out of the data collection and when we merge the Neutral with the Non-Credible class. Furthermore we run experiments with different levels of confidence for our ground truth data set of both category and credibility. We expect to see an improvement in performance if we only use the tweets with a higher confidence level while at the same time reducing the size of our data set. In our experiments we will investigate what a good tradeoff is regarding the confidence level.

We use a k-fold cross validation strategy with k = 10 which separates the data into 10 equal size subsets and which uses a single subset as test data and the other nine subsets as training data for the classifier. This process is done 10 times where each time a different subset is used as test data. The reason we use k-crossfold validation is to reduce the variance in the performance estimate of the training and test set since our data set is small. We choose k = 10 similar to Kang et al.[23] because it gives a good tradeoff between the variance of the estimate and the computation time as well as being a common value for k.

We also investigate different subsets of features with the CFS (Correlation feature

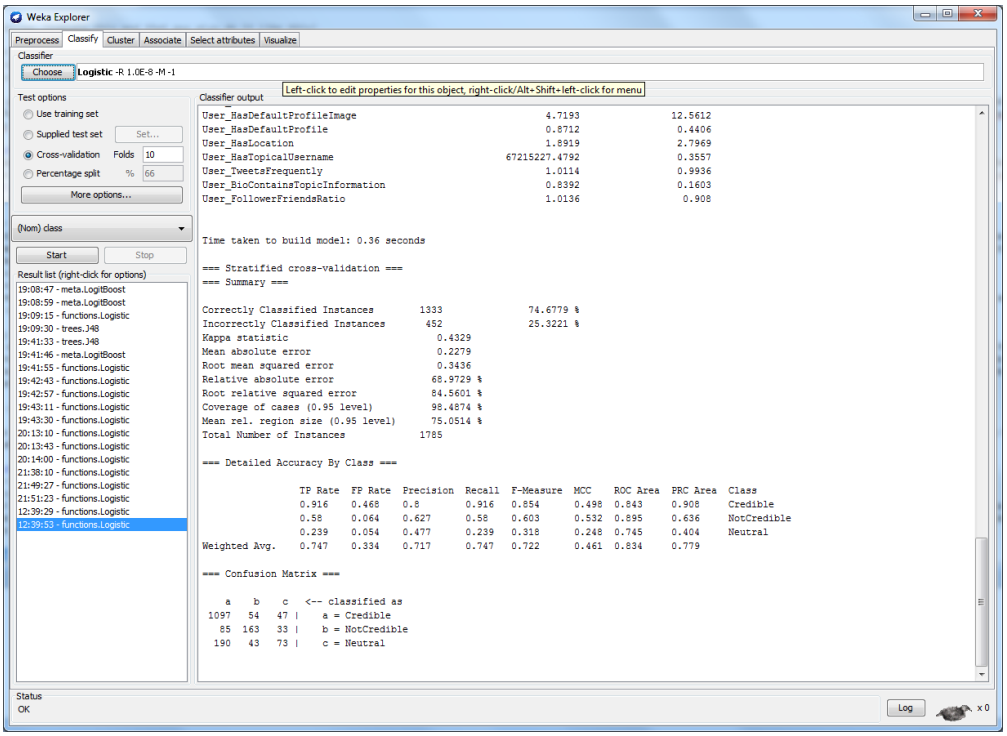



Figure 4.5: Graphical User Interface of WEKA

selection) evaluation method using a best first strategy. CFS evaluates which subset of features have a high correlation with the class while having a low correlation between each other. By using CFS we can see which features are most important with regards to credibility and category.

### 4.4.2 WEKA

For our research we use the tool WEKA[6] developed by the University of Waikato in New Zealand. WEKA is an open source software program written in Java that is used for data analysis and data mining. The program contains several tools which are used for classification, clustering, association and visualization. We use WEKA for pre-processing our data, running the classification tasks as well as finding the best feature subset. The program contains a simple GUI (Graphical User Interface) as shown in Figure 4.5 that allows us to quickly see what the results are for our research.

# Chapter 5

# Credibility evaluation

In the previous chapter we described how we collected our data and created a ground truth data set for credibility and category. We also outlined the research methods and tools which we will use for our experiments. In this chapter we describe the results of our research for determining the category and the credibility of tweets. We first investigate classifying the category since we will use the category as a feature for the credibility classification. We evaluate the use of different confidence levels as well as different classifiers for the best performance. Next we do a feature analysis to see which feature set is most important in determining the category. Second we investigate classifying the credibility of cancer tweets. Just as with the category research we will also evaluate the different confidence levels and classifiers to find the best performance. Furthermore we will also look at the effect of the Neutral class on the results and we experiment by leaving out this class and merging it together with the Non-Credible class. Next we analyze our data and state the findings of our research.

## 5.1 Determining category

In this section we investigate how we can assign a category to a tweet by using a classifier. As indicated we have 5 different categories into which we can divide the tweets which are: Advice, News, Event/Fundraise, Personal and Other. We use the ground truth data set that we created in section 4.2 which contains 999 tweets with a division as shown in Table 5.1.

We now experiment with different classifiers to determine which one performs the best for assigning a category to a tweet. We investigate the performance of Logit boost, Logistic, Bayesian classifiers as well as a range of other classifiers. We use our

Table 5.1: Category data set division

| Category | Amount | Percentage |
|---|---|---|
| News | 362 | 36% |
| Advice | 112 | 11% |
| Event/Fundraise | 191 | 19% |
| Personal | 275 | 28% |
| Other | 59 | 6% |

Table 5.2: Classifier results for Category

| Logistic - Accuracy 61.76% | | | | | |
|---|---|---|---|---|---|
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Advice | 0.038 | 0.346 | 0.161 | 0.22 | 0.668 |
| Event/Fundraise | 0.075 | 0.573 | 0.429 | 0.491 | 0.794 |
| News | 0.235 | 0.676 | 0.865 | 0.759 | 0.865 |
| Other | 0.012 | 0.154 | 0.034 | 0.056 | 0.59 |
| Personal | 0.174 | 0.616 | 0.735 | 0.67 | 0.863 |
| Weigth. Avg. | 0.153 | 0.572 | 0.618 | 0.581 | 0.813 |
| MultiClass - Accuracy 61.76% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Advice | 0.028 | 0.39 | 0.143 | 0.209 | 0.674 |
| Event/Fundraise | 0.073 | 0.563 | 0.398 | 0.466 | 0.79 |
| News | 0.248 | 0.667 | 0.876 | 0.757 | 0.867 |
| Other | 0.007 | 0.125 | 0.017 | 0.03 | 0.617 |
| Personal | 0.184 | 0.609 | 0.753 | 0.673 | 0.864 |
| Weigth. Avg. | 0.158 | 0.568 | 0.618 | 0.574 | 0.815 |
| BayesNet - Accuracy 61.66% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Advice | 0.028 | 0.375 | 0.134 | 0.197 | 0.645 |
| Event/Fundraise | 0.111 | 0.508 | 0.487 | 0.497 | 0.794 |
| News | 0.209 | 0.693 | 0.829 | 0.755 | 0.871 |
| Other | 0.012 | 0.313 | 0.085 | 0.133 | 0.584 |
| Personal | 0.171 | 0.621 | 0.738 | 0.674 | 0.867 |
| Weigth. Avg. | 0.148 | 0.58 | 0.617 | 0.584 | 0.813 |

data set of 999 tweets as stated above and use 10-fold cross validation. For our initial tests with the different classifiers we do not leave out tweets based on the confidence level. The best results for assigning a tweet to a certain category were obtained with the Logistic, MultiClass and BayesNet classifiers for which the results are shown in Table 5.2. All three of these classifiers achieve about the same accuracy of 62%. If we look at the F-Measure we see that the results are average for the Event/Fundraise, News and Personal categories and poor results for the Advice and Other category. The poor results for the Other category is be explained by the fact that the Other category functions as a rest category for tweets that do not fit into any of the first four categories and the Other category is therefore more diverse in terms of features that describe this class.

We now test the difference in performance for these three classifiers with different confidence levels. As we explained in Section 4.2 the confidence level for each labeled tweet indicates the agreement score of the people who labeled the tweets. A higher score indicates more certainty about the label of a tweet. Here we make a tradeoff between improving the performance of the classifiers and over-training the classifiers with a smaller data set as seen in Table 5.3. After several initial experiments we have decided to compare the following confidence levels: 0, 0.5 and 0.6. The performance results of our classification experiment with confidence levels 0.5 is shown in Table

Table 5.3: Amount of tweets per confidence level

| | |
|---|---|
| All | 999 |
| 0.5 | 895 |
| 0.6 | 773 |

Table 5.4: Category classification results for confidence level 0.5

| Logistic - Accuracy 64.69% | | | | | |
|---|---|---|---|---|---|
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Advice | 0.026 | 0.432 | 0.178 | 0.252 | 0.676 |
| Event/Fundraise | 0.087 | 0.589 | 0.497 | 0.539 | 0.804 |
| News | 0.208 | 0.709 | 0.858 | 0.777 | 0.873 |
| Other | 0.026 | 0.083 | 0.048 | 0.061 | 0.606 |
| Personal | 0.146 | 0.665 | 0.742 | 0.702 | 0.873 |
| Weigth. Avg. | 0.139 | 0.616 | 0.647 | 0.622 | 0.827 |
| MultiClass - Accuracy 65.03% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Advice | 0.021 | 0.485 | 0.178 | 0.26 | 0.681 |
| Event/Fundraise | 0.078 | 0.594 | 0.458 | 0.517 | 0.804 |
| News | 0.218 | 0.703 | 0.877 | 0.78 | 0.883 |
| Other | 0.014 | 0.143 | 0.048 | 0.071 | 0.656 |
| Personal | 0.163 | 0.645 | 0.758 | 0.697 | 0.881 |
| Weigth. Avg. | 0.145 | 0.617 | 0.65 | 0.619 | 0.836 |
| BayesNet - Accuracy 66.03% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Advice | 0.03 | 0.314 | 0.122 | 0.176 | 0.646 |
| Event/Fundraise | 0.106 | 0.561 | 0.542 | 0.551 | 0.811 |
| News | 0.178 | 0.74 | 0.858 | 0.795 | 0.895 |
| Other | 0.014 | 0.25 | 0.095 | 0.138 | 0.61 |
| Personal | 0.143 | 0.678 | 0.77 | 0.721 | 0.892 |
| Weigth. Avg. | 0.131 | 0.621 | 0.66 | 0.632 | 0.839 |

5.4 and with confidence level 0.6 is shown in Table 5.5. All classifiers perform better with a higher confidence level which can be explained by the fact that tweets with a higher confidence level are easier to assign to a certain category since there is less ambiguity about them. If we look at the 0.5 confidence level, there's about a 3% to 4.5% increase in terms of accuracy. Overall there are also slightly higher F-Measures for all categories which indicates a performance improvement.
For the 0.6 confidence level there's about a 5.5% to 6.5% increase in terms of accuracy. Once again we can see that overall there are also slightly higher F-Measures for all categories. Higher confidence levels improve the performance of our classifiers even more, but results in our data set becoming smaller than used by Kang et al.[23] which we use as a minimum size for our data set since we have 5 classes as opposed to 4. Based on our results we will use the confidence level 0.6 since it gives the best tradeoff between the size of our data set and the performance.

Table 5.5: Category classification results for confidence level 0.6

| Logistic - Accuracy 68.31% | | | | | |
|---|---|---|---|---|---|
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Advice | 0.031 | 0.405 | 0.214 | 0.28 | 0.693 |
| Event/Fundraise | 0.079 | 0.622 | 0.479 | 0.541 | 0.771 |
| News | 0.186 | 0.751 | 0.912 | 0.823 | 0.899 |
| Other | 0.016 | 0.2 | 0.1 | 0.133 | 0.653 |
| Personal | 0.132 | 0.688 | 0.762 | 0.723 | 0.884 |
| Weigth. Avg. | 0.128 | 0.653 | 0.683 | 0.659 | 0.839 |
| MultiClass - Accuracy 67.66% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Advice | 0.03 | 0.4 | 0.2 | 0.267 | 0.692 |
| Event/Fundraise | 0.071 | 0.626 | 0.436 | 0.514 | 0.761 |
| News | 0.209 | 0.73 | 0.918 | 0.813 | 0.908 |
| Other | 0.009 | 0.222 | 0.067 | 0.103 | 0.661 |
| Personal | 0.141 | 0.676 | 0.771 | 0.721 | 0.888 |
| Weigth. Avg. | 0.137 | 0.643 | 0.677 | 0.647 | 0.842 |
| BayesNet - Accuracy 67.14% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Advice | 0.02 | 0.462 | 0.171 | 0.25 | 0.643 |
| Event/Fundraise | 0.095 | 0.6 | 0.527 | 0.561 | 0.805 |
| News | 0.175 | 0.753 | 0.871 | 0.808 | 0.902 |
| Other | 0.016 | 0 | 0 | 0 | 0.494 |
| Personal | 0.154 | 0.656 | 0.766 | 0.707 | 0.893 |
| Weigth. Avg. | 0.132 | 0.638 | 0.671 | 0.645 | 0.84 |

## 5.2 Feature analysis category

Next we do an analysis of the features to see which feature subset best distinguishes the tweets between the different categories. We use a CFS evaluation strategy (Correlation feature selection) with a best first strategy as explained in section 4.4.1. We run the CFS evaluation on the data set with a confidence of 0.6. In total 430 subsets were evaluated with WEKA with the best subset shown below with a merit of 0.277 with 8 features. Looking at these results we see that the the "'Number of Urls"' is important which is explained due to the fact that lots of News tweets contains links and are therefore more easily identifiable.

- Number of URLs
- Has exclamation mark
- Has pronoun first
- Has pronoun second
- Number of intensifiers
- Is a reply

Table 5.6: Credibility data set division

| | | |
|---|---|---|
| Credible | 722 | 58% |
| Neutral | 276 | 22% |
| Non-Credible | 195 | 16% |

- User tweets frequently
- Favorites count of the user

## 5.3 Determining credibility

We now investigate how to automatically assign a cancer tweet to a credibility level. We have 3 classes for credibility as mentioned in section 4.2 which are: Credible, Neutral and Non-Credible. The class "'Can't decide"' is not used as there are too few instances in our ground truth data set. The data set we use, leaving out the "'Can't decide"' class consists of 1238 tweets and a division as shown in Table 5.6

Just as with the category research we experiment with different classifiers to determine which one performs the best for assigning a credibility to a tweet. We investigate the performance of the J48 decision tree, to compare our results with Castillo et al.[11], and Logistic, Bayesian classifiers as well as a range of other classifiers. We use 10-fold cross-validation and perform the experiments with WEKA. The best results were obtained with the ClassificationViaRegression and LogitBoost classifiers which achieved around 65% accuracy in terms of performance. The J48 decision tree performed worse with an accuracy of only 60% as shown in Table 5.7. The F-Measures shows that we achieve a good performance for the Credible class, poor performance for the Neutral class and very poor performance for the Non-Credible class with all three classifiers. A reason for the poor performance for the Neutral class is that there is less agreement about whether or not a tweet is Neutral and thus it does not have very good distinguishing features. We will investigate this further when we experiment with different confidence levels. An explanation of the poor performance for the Non-Credible class is the fact that our data set contains only 16% tweets which are labeled as Non-Credible. From the confusion matrix we can see that the Neutral class causes problems with a high number of Neutral items being classified as Credible and items that are Non-Credible as Neutral. We will therefore investigate what the effect on performance is when we merge the Neutral and Non-Credible class and when we remove the Neutral class.

We first look at the effect of merging the Neutral and the Non-Credible class similar to Castillo et al.[12]. We once again experiment with a range of different classifiers to find the best performance. The performance results of the best 10 classifiers do not differ much and we will therefore use the same three classifiers as in the previous experiment in order to compare the results. We can see from Table 5.8 that the accuracy improves with about 7-8% and the F-Measures has significantly improved for the Non-Credible class with an average results of around 0.6. The improved results can be explained by the fact that we reduced the problem to that of a binary classification which is simpler that multiclass classification.

Table 5.7: Classifier results for credibility with Credible, Neutral and Non-credible classes

| ClassificationViaRegression - Accuracy 63.54% | | | | | |
|---|---|---|---|---|---|
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.692 | 0.673 | 0.928 | 0.78 | 0.744 |
| Non-Credible | 0.064 | 0.462 | 0.282 | 0.35 | 0.748 |
| Neutral | 0.049 | 0.423 | 0.12 | 0.186 | 0.623 |
| Weight. avg. | 0.441 | 0.581 | 0.635 | 0.572 | 0.717 |
| LogitBoost - Accuracy 63.87% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.563 | 0.71 | 0.898 | 0.793 | 0.743 |
| Non-Credible | 0.064 | 0.453 | 0.272 | 0.34 | 0.751 |
| Neutral | 0.111 | 0.374 | 0.221 | 0.278 | 0.604 |
| Weight. avg. | 0.377 | 0.59 | 0.639 | 0.6 | 0.307 |
| J48 - Accuracy 59.26% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.548 | 0.692 | 0.802 | 0.743 | 0.601 |
| Non-Credible | 0.095 | 0.438 | 0.379 | 0.407 | 0.621 |
| Neutral | 0.145 | 0.289 | 0.196 | 0.233 | 0.467 |
| Weight. avg. | 0.381 | 0.557 | 0.593 | 0.57 | 0.573 |

Table 5.8: Classifier results for credibility with Credible and Non-Credible classes

| ClassificationViaRegression - Accuracy 69.74% | | | | | |
|---|---|---|---|---|---|
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.452 | 0.729 | 0.795 | 0.761 | 0.742 |
| Non-Credible | 0.205 | 0.635 | 0.548 | 0.588 | 0.742 |
| Weight. avg. | 0.355 | 0.692 | 0.697 | 0.693 | 0.742 |
| LogitBoost - Accuracy 72.09% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.423 | 0.747 | 0.814 | 0.779 | 0.742 |
| Non-Credible | 0.186 | 0.67 | 0.577 | 0.62 | 0.742 |
| Weight. avg. | 0.329 | 0.717 | 0.721 | 0.717 | 0.742 |
| J48 - Accuracy 67.48% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.456 | 0.719 | 0.76 | 0.739 | 0.641 |
| Non-Credible | 0.24 | 0.597 | 0.544 | 0.569 | 0.641 |
| Weight. avg. | 0.371 | 0.67 | 0.675 | 0.672 | 0.641 |

Table 5.9: Classifier results for credibility with Credible and Non-Credible classes and Neutral removed

| ClassificationViaRegression - Accuracy 81.90% | | | | | |
|---|---|---|---|---|---|
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.636 | 0.846 | 0.942 | 0.891 | 0.792 |
| Non-Credible | 0.058 | 0.628 | 0.364 | 0.461 | 0.792 |
| Weight. avg. | 0.513 | 0.8 | 0.819 | 0.8 | 0.792 |
| LogitBoost - Accuracy 82.44% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.615 | 0.85 | 0.943 | 0.894 | 0.803 |
| Non-Credible | 0.057 | 0.647 | 0.385 | 0.482 | 0.803 |
| Weight. avg. | 0.497 | 0.807 | 0.824 | 0.807 | 0.803 |
| J48 - Accuracy 81.13% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.472 | 0.874 | 0.888 | 0.881 | 0.687 |
| Non-Credible | 0.112 | 0.56 | 0.528 | 0.544 | 0.687 |
| Weight. avg. | 0.395 | 0.808 | 0.811 | 0.809 | 0.687 |

Next we look at the effect of removing the Neutral class from the data set in terms of the performance. Similar to the experiment with merging the Neutral and the Non-Credible class we test a range of different classifiers. Once again the performance results of the best 10 classifiers do not differ much and we will therefore use the same three classifiers as in our previous two credibility experiments so we can compare the results. We can see from Table 5.9 that the accuracy improves significantly with about 21% compared to the original data set with three classes. The increase in performance is due to the problem being reduced to a binary classification from a multiclass classification. Furthermore the improvement in the results from removing the Neutral class compared to merging it is because without the tweets of the Neutral class there is a better distinction in the ground truth between the Credible and Non-credible classes. The best performance is achieved by the Logit Boost classifier with a 82.44% accuracy and a 0.807 F-Measure.

From the previous experiments we can see that merging the Neutral class with the Non-credible class as well as completely removing the Neutral class items from the data set give an increase in performance. However we decide to continue our research with the original three classes (Credible, Neutral and Non-credible) as we did in our original setup. Leaving out the Neutral class would give us a data set that is skewed and no longer a good reflection of the actual cancer tweets on Twitter. Merging the Neutral and the Non-credible class reduces our problem to a binary classification but this simplication is not supported by the labeling of the workers on CrowdFlower. The workers labeled part of our data set as Neutral instead of Non-credible and they thus believe there is a difference between these two classes. Finally, just a with our category experiment we will see what the effect of only using tweets with a minimum confidence score has on our results. We choose a confidence level of 0.5 as the minimum as higher confidence levels makes the data set too small. Using a 0.5 confidence interval leaves us with a data set of 734 tweets which we will use to test the three classifiers of the previous experiments. We can see from Table 5.10 that the best result is

Table 5.10: Classifier results for credibility with Credible Neutral, Non-Credible classes with 0.5 confidence

| ClassificationViaRegression - Accuracy 71.80% | | | | | |
|---|---|---|---|---|---|
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.584 | 0.786 | 0.924 | 0.849 | 0.797 |
| Non-Credible | 0.037 | 0.39 | 0.239 | 0.296 | 0.872 |
| Neutral | 0.091 | 0.411 | 0.24 | 0.303 | 0.684 |
| Weight. avg. | 0.431 | 0.671 | 0.718 | 0.684 | 0.78 |
| LogitBoost - Accuracy 74.52% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.557 | 0.795 | 0.93 | 0.857 | 0.81 |
| Non-Credible | 0.025 | 0.564 | 0.328 | 0.415 | 0.895 |
| Neutral | 0.081 | 0.505 | 0.312 | 0.386 | 0.709 |
| Weight. avg. | 0.408 | 0.713 | 0.745 | 0.718 | 0.797 |
| J48 - Accuracy 69.21% | | | | | |
| | FP Rate | Precision | Recall | F-Measure | ROC Area |
| Credible | 0.584 | 0.778 | 0.879 | 0.825 | 0.622 |
| Non-Credible | 0.055 | 0.383 | 0.343 | 0.362 | 0.563 |
| Neutral | 0.103 | 0.362 | 0.221 | 0.274 | 0.497 |
| Weight. avg. | 0.435 | 0.654 | 0.692 | 0.667 | 0.59 |

obtained by the Logit Boost classifier with an accuracy of 74.52% and an F-Measure of 0.718. The Logit Boost classifier achieves a Kappa statistic of 0.3577 which indicates a relatively fair to poor performance over a random predictor.

## 5.4 Feature analysis credibility

Next we do an analysis of the features to see which feature subsets best distinguish the tweets between the different credibility levels. We use a CFS evaluation strategy (Correlation feature selection) with a best first strategy as explained in section 4.4.1. We run the CFS evaluation on the data set with a confidence of 0.5 with all 3 classes. In total 371 subsets were evaluated with WEKA with the best subset shown below with a merit of 0.169 with 7 features. We will further investigate the fact that a tweet containing url is important in distinguishing the credibility levels when we evaluate the comments.

- Number of Urls
- Number of upper case characters
- Number of swear words
- Tweet was posted during day time
- Followers amount of User
- Favorites amount of User

- Category

## 5.5 Data analysis

### 5.5.1 Relation between category and credibility

We now analyze the ground truth data set that we have created and look at the relationship between the credibility level and the category. In Table 5.11 we show the amount of tweets of a certain credibility level per category. We can see that the News and Personal categories contain the largest percentage of Credible tweets of the five categories. The Other category has the largest percentage of Non-Credible tweets followed by the Personal and Advice categories. The tweets in the Event and News categories have a lowest percentage of Non-Credible labels. For the Neutral credibility level we see that the Personal category is most often seen as Neutral. Looking at the comments in section 5.5.2 we can see that this is because people feel they can't judge tweets that are Personal and therefore choose the Neutral option.

Table 5.11: Category vs Credibility

| | All | | | | |
|---|---|---|---|---|---|
| | Advice | Event | News | Personal | Other |
| Credible | 62% | 75% | **79%** | 42% | 41% |
| Non-Credible | 12% | 2% | 5% | 20% | 31% |
| Neutral | 26% | 23% | 16% | 38% | 28% |
| | 0.5 confidence | | | | |
| | Advice | Event | News | Personal | Other |
| Credible | 71% | 79% | **85%** | 43% | 48% |
| Non-Credible | 13% | 1% | 2% | 19% | 35% |
| Neutral | 16% | 20% | 13% | 38% | 17% |
| | 0.6 confidence | | | | |
| | Advice | Event | News | Personal | Other |
| Credible | 75% | **85%** | **86%** | 42% | 49% |
| Non-Credible | 11% | 0% | 2% | 20% | 38% |
| Neutral | 14% | 15% | 12% | 38% | 14% |

### 5.5.2 Evaluating comments

For our labeling tasks on CrowdFlower we have asked people to support their choice for a certain category or credibility with a short comment. In this section we describe what our research of the comments of the CrowdFlower workers has yielded and the conclusions that we can draw from this. We make a distinction between the explanations of workers for a category and the explanations for a credibility level. We use the term explanation and comment interchangeably.

### Category

For the category task the CrowdFlower workers had a choice between 5 different categories: News, Advice, Event/Fundraise, Personal and Other. Every time the worker picked a category for a tweet we also asked them to explain their decision in a few words. With these explanations we can better understand what affects the choice for a certain category. In total we have 5763 explanations from workers on CrowdFlower.

**General** For all categories we found that the explanations contained part of the tweet content indicating the users view that part of the tweet as the reason for picking that category. However since our CrowdFlower task is a low paid task users could have also simply copy pasted part of the tweet content in the explanation text box to save time. Another discovery is that a small part of the explanations, 175 rows, only contained the category which the user chose and does therefore not contain any useful information. From the other explanations we can see that the textual content of the tweet is the most important part of the tweet for choosing a category. This is consistent with our feature analysis for the category evaluation in section 5.2 in which we saw that 5 of the 8 important features of the tweet are textual.

**Advice** For the tweets which were given the label Advice the workers indicated that tweets containing action, information or statements were an indicator for them to label the tweet as advice. A second comment by workers was that tweets which contained links to websites giving advice were indicative of Advice tweets.

**Event** For the Event category the workers indicated that specific keywords related to the Event category such as event, fundraiser, donation and money were important for them to choose Event for the tweet.

**News** The fact that tweets contained a link was a main reason for people, in 270 cases, to label these tweets as News. Another indicator for News was text containing research announcements. People commented that specific keywords such as study made them classify the tweet as News

**Personal** Tweet content that referenced people such as father, daughter, sister, mother, friend or their names were given as a reason for people to classify a tweet as the category Personal. A sub set of this were tweets in which people wished others well.

**Other** The Other category is a category for tweets who do not fit any of the other categories. Workers also mentioned this in the comments that they picked Other because the tweet did not fit any of the other categories. In the comments workers mentioned that the tweet was a joke or a job advert as examples of reasons to pick Other.

### Credibility

Similar to our category task we asked our workers on CrowdFlower to provide an explanation when choosing a credibility level for our credibility task. For the credibility

task workers had four choices to pick from: Credible, Neutral, Non-credible and I can't decide. In total we have 6839 explanations from workers on CrowdFlower.

**General** For all credibility levels we see that the tweet content is the most cited reason in the comments by the workers for choosing a particular credibility level which was the same with the labeling of the category. As we saw with comments from the category labeling task, part of the comments do not contain any useful information. Workers simply stated the credibility level they picked or did not fill in any useful information regarding the reason for their choice. None of the workers mentioned the profile image of the user as reason for their choice.

**Credible** Workers cited links in a tweet as a factor for them to view the tweet as credible which follows our findings of the credibility feature analysis in section 5.4 where the number of URLs is one of the top 6 best features to determine the credibility level. Other explanations mentioned many followers and the fact that the tweet contained a personal story as reasons for viewing the tweet as credible.

**Neutral** There were 111 mentions of followers as reason for picking the Neutral label by the CrowdFlower workers. The fact that a tweet was seen as personal was a reason for workers to label it as Neutral. From this we can see that workers view personal tweets as Credible or Neutral but not Non-credible.

**Non-credible** The lack of a link, or as stated by the workers the fact that no source was provided made the tweet seem Non-credible. Few followers and the fact that a tweet was a joke were also provided as reasons for labeling a tweet as Non-credible.

# Chapter 6

# Conclusions and Future Work

This chapter gives an overview of the project's contributions. First we present our conclusion and our contributions. Next we discuss possible future work which we believe is important to research next.

## 6.1 Conclusions

An increasing number of users is using social media to discuss as well as look up information about health. A lot of these users have little online experience with the Web and have difficulty assessing the credibility of the information which they find. Anyone on Twitter can post tweets about health no matter what their medical background is. We want to use automated methods to help users assess the credibility of health information on Twitter. By providing users with a supporting assessment they will be better equipped to determine the credibility of health tweets. We will now answer our main research questions as posed in the introduction. Can we support users in making credibility assessments of health tweets?

To answer this question we divided our research into a number of sub problems with a research question for each part. First we investigated whether our perceived credibility problem actually exists. We did an initial investigation in Chapter 3 in which we collected a small data set of cancer tweets from Twitter via the API using a small set of keywords related to cancer. We manually went through a subset of the data set and found that according to the authors opinion there were tweets which we considered not credible. To verify our findings we set up a small task on CrowdFlower and asked workers to label the tweets with a credibility level. From this CrowdFlower experiment we concluded that there are tweets about cancer which people view as not credible. For our main research project we defined three credibility levels to which we would then assign tweets: Credible, Neutral and Non-credible. We also noticed that there are different types of health tweets and we therefore also investigated the impact of the category of tweet on the credibility by using five categories: Advice, Event/Fundraise, News, Personal and Other. Next we created a ground truth data set in Chapter 4 by first collecting a larger data set of tweets from Twitter via their Streaming API. To find the relevant tweets we used a manually created list of keywords from the Cancer.gov[2] website. Next we set up a larger CrowdFlower experiment to create a labeled data set which we use as a ground truth for our research. To ensure the

accuracy and validity of the data we have used existing research for the design of our CrowdFlower experiments. We then compiled a list of features of our set of tweets that we use in the process of determining the credibility and category of a tweet. Then in Chapter 5 we started off by using and evaluating classifiers to automatically assign a category to a tweet. Our results show that the classifiers perform reasonably well for the categories Event/Fundraise, News and Personal but quite poor for the Advice and Other categories. We achieved an accuracy of 68% with the Logistic classifier. Next we trained and evaluated a set of classifiers to automatically assign a credibility level to a tweet. We compared the results of using all three credibility levels Credible, Neutral and Not-credible, merging the Neutral and Non-credible class and removing the Neutral class altogether. The best result was obtained by removing the Neutral class with the Logit Boost classifier with 82% accuracy. For our final result we choose to use all three classes since it best represents the division of tweets on Twitter and because workers on CrowdFlower indicated that there was a difference between the Neutral and Non-Credible classes. We achieved our best result with all three classes with the Logit boost classifier with an accuracy of 76%. From these results we conclude that we can support the user in making credibility assessments of health tweets. However the performance for detecting Non-credible tweets when using three credibility levels is still below average. While we have shown that assisting users in the health domain is possible more work is needed to improve these results before they are ready for actual use as we will describe next in Future work.

## 6.2 Future work

There are a number of follow up research steps which we feel are important for this research. First and foremost we believe that our research should be used to create a prototype application to assist the user in making credibility assessments of health tweets. User surveys should be used to collect important feedback and to make improvements to the methods used for the credibility assessments. Another next step would be to automate the process of collecting keywords for obtaining a data set from Twitter by using existing health databases to create lists of keywords. Furthermore we believe that we can further improve the results by investigating the difference in credibility perception among users. While the scope of our research was limited to the disease cancer, we believe the methods used in this research can be applied to the general health domain with some modifications. Finally we want to investigate the creation of a credibility score where we use the output of the classifiers to generate a credibility score for each tweet with a simple explanation for the user telling them what the score is based on.

# Bibliography


[1] Cancer facts figures 2014. http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/index. Accessed: 2013-10-30.

[2] Comprehensive cancer information - national cancer institute. http://www.cancer.gov/. Accessed: 2012-08-30.

[3] Merriam-webster dictionary cancer. http://www.merriam-webster.com/dictionary/cancer. Accessed: 2012-08-30.

[4] Merriam-webster dictionary credible. http://www.merriam-webster.com/dictionary/credible. Accessed: 2013-10-30.

[5] Twitter company statistics. https://about.twitter.com/company. Accessed: 2014-07-01.

[6] Weka. http://www.cs.waikato.ac.nz/ml/weka/. Accessed: 2012-08-30.

[7] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. Semantics+ filtering+ search= twitcident. exploring information in social web streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 285–294. ACM, 2012.

[8] Omar Alonso, Chad Carson, David Gerster, Xiang Ji, and Shubha U Nabar. Detecting uninteresting content in text streams. In *SIGIR Crowdsourcing for Search Evaluation Workshop*, 2010.

[9] Omar Alonso and Stefano Mizzaro. Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, volume 15, page 16, 2009.

[10] Kevin Robert Canini, Bongwon Suh, and Peter L Pirolli. Finding credible information sources in social networks based on content and social structure. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 1–8. IEEE, 2011.

[11] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM.

[12] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588, 2013.

[13] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2451–2460. ACM, 2012.

[14] Ernesto Diaz-Aviles, Avaré Stewart, Edward Velasco, Kerstin Denecke, and Wolfgang Nejdl. Epidemic intelligence for the crowd, by the crowd. In *ICWSM*, 2012.

[15] BJ Fogg and Hsiang Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 80–87. ACM, 1999.

[16] Sandra Garcia Esparza, Michael P O'Mahony, and Barry Smyth. Catstream: categorising tweets for user profiling and stream filtering. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 25–36. ACM, 2013.

[17] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 575–590. ACM, 2012.

[18] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 729–736. International World Wide Web Conferences Steering Committee, 2013.

[19] Carleen Hawn. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*, 28(2):361–368, 2009.

[20] Panagiotis G Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.

[21] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.

[22] Sue Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. I can't get no sleep: discussing# insomnia on twitter. In *Proceedings of*

*the SIGCHI Conference on Human Factors in Computing Systems*, pages 1501–1510. ACM, 2012.

[23] Byungkyu Kang, John O'Donovan, and Tobias Höllerer. Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, pages 179–188, New York, NY, USA, 2012. ACM.

[24] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2583–2586. ACM, 2012.

[25] MichałKąkol, MichałJankowski-Lorek, Katarzyna Abramczuk, Adam Wierzbicki, and Michele Catasta. On the subjectivity and bias of web content credibility evaluations. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 1131–1136, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[26] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, New York, NY, USA, 2012. ACM.

[27] Dhiraj Murthy, Alexander Gross, and Daniela Oliveira. Understanding cancer-based networks in twitter using social network analysis. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 559–566. IEEE, 2011.

[28] John O'Donovan, Byungkyu Kang, Greg Meyer, Tobias Hollerer, and Sibel Adalii. Credibility in context: An analysis of feature distributions in twitter. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 293–301, Washington, DC, USA, 2012. IEEE Computer Society.

[29] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, pages 265–272, 2011.

[30] Michael J Paul and Mark Dredze. A model for mining public health topics from twitter. *HEALTH*, 11:16–6, 2012.

[31] Maria Rafalak, Katarzyna Abramczuk, and Adam Wierzbicki. Incredible: Is (almost) all web content trustworthy? analysis of psychological factors related to website credibility evaluation. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 1117–1122, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[32] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[33] Julia Schwarz and Meredith Morris. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1245–1254. ACM, 2011.

[34] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[35] S John Sullivan, Anthony G Schneiders, Choon-Wi Cheang, Emma Kitto, Hopin Lee, Jason Redhead, Sarah Ward, Osman H Ahmed, and Paul R McCrory. âwhat's happening?âa content analysis of concussion-related traffic on twitter. *British journal of sports medicine*, 46(4):258–263, 2012.

[36] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.

# Appendix A

# Keywords

Below is given an overview of the cancer related keywords used to collect our data set for this research project.
acinic cell carcinoma
acoustic neuroma
acral lentiginous melanoma
acrospiroma
adamantinoma
adenocarcinoma
adenoid cystic carcinoma
adenoma
adenomatoid odontogenic tumor
adult t-cell leukemia
aggressive nk-cell leukemia
aldesleukin
ameloblastic fibroma
anal cancer
angiomyolipoma
appendix cancer
arsenic trioxide
asbestos
asparaginase
astrocytoma
astrocytomas
atypical teratoid rhabdoid tumor
b-cell leukemia
b-cell lymphoma
basal cell carcinoma
beacopp
bevacizumab
bexarotene
bile duct cancer
biopsy cancer
bladder cancer

blastoma
bleomycin
bone cancer
bone marrow transplant
bone tumor
bortezomib
bowen's disease
brachytherapy
brain stem glioma
brain tumor
breast cancer
brenner tumour
bronchial tumor
bronchioloalveolar carcinoma
brown tumor
burkitt lymphoma
busulfan
cancer
cancer causes
cancer immunotherapy
cancer tumor
capecitabine
carboplatin
carcinogen(s)
carcinoid
carcinoid tumor
carcinoma
carcinoma in situ
carcinoma of the penis
carcinoma of unknown primary
carcinosarcoma
carmustine
cavernous angioma
cervical cancer
cervical conization
chemotherapy
chlorambucil
cholangiocarcinoma
chondroma
chondrosarcoma
choriocarcinoma
choroid plexus tumor
cisplatin
cladribine
colon cancer
colorectal cancer
copp-abvd regimen

corticosteroid
cowden syndrome
craniopharyngioma
crizotinib
cromoglicic acid
cryoablation
cryosurgery
cutaneous t-cell lymphoma
cyclophosphamide
cytarabine
dacarbazine
dactinomycin
daunorubicin
denileukin diftitox
dermatofibrosarcoma protuberans
dermoid cyst
desmoplastic small round cell tumor
diethylstilbestrol
disseminated malignant neoplasm
doctor cancer
doxorubicin
duct bile cancer
ductal carcinoma in situ
dysplastic nevus
embolization
embryonal carcinoma
embryonal tumor
endodermal sinus tumor
endometrial cancer
endometrial cancer
endometrioid tumor
enucleation
ependymoma
epithelial cancer
epithelioid sarcoma
erlotinib
erythroleukemia
esophageal cancer
esophagectomy
esophagogastrectomy
esthesioneuroblastoma
etoposide
ewing sarcoma
external beam radiotherapy
extragonadal germ cell tumor
eye cancer
fallopian tube cancer

familial adenomatous polyposis
fast neutron therapy
fetus in fetu
fibroma
fibrosarcoma
fluorouracil
follicular lymphoma
follicular thyroid cancer
fosfestrol
gallbladder cancer
ganglioneuroma
gardner's syndrome
gastric cancer
gastric lymphoma
gastrointestinal cancer
gastrointestinal carcinoid tumor
gastrointestinal stromal tumor
gastrointestinal tumor
germ cell tumor
germinoma
gestational choriocarcinoma
gestational trophoblastic tumor
giant cell tumor of bone
glioblastoma multiforme
glioma
gliomatosis cerebri
gliosarcoma
glomus tumor
glucagonoma
gonadoblastoma
granulocytic sarcoma
granulosa cell tumour
hairy cell leukemia
head cancer
heart cancer
hemangioblastoma
hemangiopericytoma
hemangiosarcoma
hematological malignancy
hematologist
hematology
hepatoblastoma
hepatocellular cancer
hepatocellular carcinoma
hepatosplenic t-cell lymphoma
hereditary nonpolyposis colorectal cancer
histiocytosis langerhans cell

hodgkin lymphoma
hodgkins lymphoma
human leucocyte-derived alpha interferons
hydroxycarbamide
hypopharyngeal cancer
ibritumomab tiuxetan
idarubicin
ifosfamide
imatinib
immunotherapy
inguinal orchiectomy
insulinoma
interferon alfa-2b
interferon alfacon-1
interferon alpha 2a
interferon beta-1a
interferon beta-1b
interferon gamma-1b
intestine cancer
intraocular melanoma
irinotecan
islet cell cancer
islet cell carcinoma
kaposi sarcoma
kidney cancer
klatskin tumor
krukenberg tumor
langerhans cell histiocytosis
laryngeal cancer
laryngectomy
larynx cancer
leiomyosarcoma
lenalidomide
leukemia
li-fraumeni syndrome
lip cancer
lipoma
liposarcoma
liposomal doxorubicin
liver cancer
liver tumour
lobectomy
lobular carcinoma in situ
lomustine
loop electrical excision procedure
lung cancer
luteoma

lymphangioma
lymphangiosarcoma
lymphoepithelioma
lymphoid leukaemia
lymphoma
malignant fibrous histiocytoma
malt lymphoma
mammogram
mammography
mantle cell lymphoma
mast cell leukemia
mastectomy
mastocytosis
mechlorethamine
medulloblastoma
medulloblastoma
melanoma
melphalan
meningioma
merkel cell cancer
merkel cell carcinoma
mesothelioma
metaplastic carcinoma
metastasis
metastasis cancer
methotrexate
metirosine
midline carcinoma
mitomycin
mitoxantrone
mohs surgery
mole cancer
monocytic leukemia
mouth cancer
mucinous tumor
multiple endocrine neoplasia
multiple myeloma
mustargen oncovin procarbazine prednisone
mycosis fungoides
myelodysplastic syndromes
myeloid leukemia
myeloproliferative neoplasms
myxoid liposarcoma
nasal cavity cancer
nasopharyngeal cancer
neck cancer
neck dissection

neoplasia
nephrectomy
nerve sheath tumor
neuroblastoma
neurofibromatosis type i
neuroma
nilutamide
nodular melanoma
non-hodgkin lymphoma
nonmelanoma
octreotide
ocular oncology
oligoastrocytoma
oligodendroglioma
oncocytoma
oncologist
oncology
optic nerve tumor
oral cancer
oropharyngeal cancer
osteosarcoma
ovarian cancer
paclitaxel
paget's disease
pancoast tumor
pancreatectomy
pancreatic cancer
pancreaticoduodenectomy
papillomatosis
paraganglioma
paranasal sinus cancer
parathyroid cancer
peginterferon alfa-2a
peginterferon alfa-2b
penile cancer
penis cancer
pharyngeal cancer
pharyngectomy
pheochromocytoma
photodynamic therapy
pinealocytoma
pituicytoma
pituitary adenoma
pituitary tumor
plasma cell neoplasm
pleuropulmonary blastoma
polyembryoma

pregnancy cancer
primary effusion lymphoma
primary peritoneal cancer
propranolol
prostate brachytherapy
prostate cancer
proteus syndrome
pseudomyxoma peritonei
radiation therapy
radiation therapy
radical perineal prostatectomy
radical retropubic prostatectomy
radioactive iodine therapy
radiosurgery
rectal cancer
renal cancer
renal carcinoma
renal cell
renal medullary carcinoma
retinoblastoma
retroperitoneal lymph node dissection
rhabdoid tumour
rhabdomyoma
rhabdomyosarcoma
richter's transformation
rituximab
sézary syndrome
sacrococcygeal teratoma
salivary gland cancer
sarcoma
schwannomas
schwannomatosis
sebaceous gland carcinoma
sertoli-leydig cell tumour
sex cord-stromal tumour
sir-spheres
skin cancer
soft tissue sarcoma
somatostatinoma
soot wart
spinal tumor
spine cancer
squamous cell carcinoma
stanford v
stem cell transplantation
stomach cancer
superficial spreading melanoma

surgery cancer
synovial sarcoma
t-cell lymphoma cutaneous
tamoxifen
temozolomide
teniposide
teratoma
testicular cancer
thalidomide
thecoma
thiotepa
throat cancer
thymic carcinoma
thymoma
thyroid cancer
thyroidectomy
tioguanine
topotecan
transitional cell carcinoma
tretinoin
tumor
tumour
urethral cancer
uterine cancer
uterine sarcoma
uveal melanoma
vaginal cancer
vemurafenib
vinblastine
vincristine
vinorelbine
vipoma
vulvar cancer
waldenström macroglobulinemia
warthin's tumor
wilms tumor
xeroderma pigmentosum
zoledronic acid