To my family

Abstract

The proliferation of Internet usage has resulted in huge amounts of information being made available online. We instinctively trust our closest friends or relatives, or information presented to us by people in our networks[2]. The fact that there is so much information, often with rumours or inconsistencies makes it hard to know who or which information source to trust, or much worse makes it hard to not trust information that we believe come from a reliable source.

This thesis looks at identifying trust metrics for online social media, particularly in the case of Twitter. We look at the characteristics of trustworthy and untrustworthy tweets for example the length of tweet, the amount of retweets, followers etc and learn patterns from this to create a model which determines if a tweet is trustworthy or not. We compare the results obtained from two different models, based on two machine learning techniques.

Acknowledgements

Contents

List of Figures

Part I

Background

Chapter 1

Introduction

In this Chapter we give you an explanation of the purpose of study and the research questions we would address. We also give the motivation and aim of the project followed by the goals that should be achieved at the end of this research and finally give a brief overview of how this thesis is structured.

1.1 Motivation

Online interactions represent a complex blend of human actors and technological systems[2]. In recent years, the type of content on the web has transformed greatly and the interactions online between human actors has become increasingly popular. A big contributor to this is social media networks such as Facebook, Twitter and Google Plus and Question and Answer platforms like Yahoo Answers, Quora and Stackoverflow which act as public boards of discussion. Opinions are frequently posted to social media sites and are often a mixture of fact, speculation or rumour whereas user-driven sites such as Wikipedia are often questioned for their trustworthiness[4]. Blogs are used as sound boards for average users to disseminate information to their social circles and favourable opinion becomes fact which in turn is regarded as trustworthy information.

Todays brands place a lot of trust in opinions on social media, use social media to learn from their audience and have integrated social media into their business model. Whilst the scale and variety of information has considerable value for commercial and social purposes, the quality, provenance, trust and validity is often questionable[1]. It is impossible for us as human beings to manually verify every piece of information presented to us on social media. It is also not possible to make an informed judgement on the trustworthiness of a person,account or entity online. Verified Twitter and Facebook profiles go some way to ensure that the identity of an entity is verified independently therefore ensuring the authenticity of that entity. However, this does not go far enough in making sure that the information on social networks is truly accurate and therefore trustworthy. Trust can be quantified through the creation of algorithms and quality metrics which can be used to address the issues of validity and quality through the use of automated assessments. These metrics would be used to identify different aspects of an informational source and look for independent verification of the quality of the data. The metrics would need to be tested against real-world examples and vary depending on the platforms (e.g : Twitter, Facebook) and the situation/context. (e.g : Movie ratings might have different metrics and weights to a crisis situation); the nuances of language can change the meaning of a piece of information drastically through the use of slightly different words. An example of this would be "*We anticipate the latest release to be announced shortly*" versus "*We expect the latest release to be announced shortly*". Although the words are synonyms, the first sentence regards the release as something that is probable. The second sentence regards the release to be more of a certainty.

Another reason why trust metrics is important is because of the huge difference in variance in user-generated content consumed by everyone today as opposed to traditional content published by news websites which was the main source of information back in the day[5]. This is particularly significant in knowledge based media such as twitter and question and answer platforms where you could have very high quality trustworthy content to low quality, untrustworthy and sometimes abusive content[5].

Trust is also a primary factor in influence and social media is now heavily relied upon as a source of information. This can vary from international breaking news to booking the next holiday. The common factor is that the information is often based on people's experiences that are then shared online. Trustworthiness then becomes of paramount importance especially considering the fact that something that becomes popular on the internet is not necessarily true and vice versa.

1.2 Aim

The aim of this work is to provide a systematic method of verifying information on the internet, specifically on twitter, independently and create a more trustworthy online experience. There has been a lot of research done in this area, which we will detail in the next section, however our approach is novel in that it specifically studies the features of news websites. Our assumption being that official news websites are trustworthy and then studies the features of some known untrustworthy twitter pages and uses this knowledge to predict the trustworthiness of tweets that it's fed. This project will help improve upon current research and look at alternative solutions than those that are proposed and provide an implementation to test these out which we believe will help brands, people and the whole of the digital world to have a better online experience.

1.3 Goals

In this project we address the task of classifying information as trustworthy or not on Twitter. We focus on the following research questions and goals :

1. Identify the factors that affect trustworthiness in Online Social media and narrow it down to the factors that could be important in the case of Twitter.

2. Data Gathering - Collect Trustworthy and Untrustworthy tweets for analysis and verification.

3. Implement a Machine Learning model and try and improve the accuracy of prediction of this model by tweaking features, adding more features that could be inferred from the features that were collected from the Twitter API, etc to give us an accuracy of 80% or more if possible.

4. Compare this with other models to pick the best model for this problem.

5. Verify the prediction of model by using new data.

1.4 Thesis Structure

Chapter 1 We first give a introduction and overview of the project, identify the goals and purpose as well as provide a motivation behind it.

Chapter 2 Provides some background into the research already done in related areas, in different domains but with similar ideas.

Chapter 3 This chapter looks specifically at Logistic Regression which was a Machine Learning Technique we used.

Chapter 4 We look at the theory behind Support Vector Machines in this Chapter which was the second Machine Learning Technique used.

Chapter 5 The method and approach to the problem is discussed here.

Chapter 7 Implementation Details are discussed.

Chapter 8 Testing and results are discussed here.

Chapter 9 Reflections, Conclusion, challenges and future work would conclude this thesis.

Chapter 2

Literature Review

Our work draws on a significant amount of research done prior to this and in this chapter we will outline and discuss some of the related work done before we introduce our model in the next chapter.

The research problem we address is identifying trustworthy content amongst data in social media. An extensive literature review was done on this topic to identify very specific articles and general ideas that could be gathered from related topics.

2.1 Research Types, Domains, Similarity

Let us first look at the different types of related research done and the domains they were carried out on and see how similar or different they are to our problem. "Finding High Quality Content in Social Media"[5] a paper by Yahoo! Research, addresses the problem of finding high quality content in community-driven question/answering sites. They focus mainly on the on question/answering domain and have achieved an accuracy close to that of humans. Quality of content doesn't necessarily relate to trustworthiness of content but from our analysis we find that they certainly share similar characteristics. One of the primary differences of this research to our problem is that the actors involved are different. Here you have an asker, answerer and an evaluator. We have just a User and a fan. For the lack of a better word, we use the word fan to describe anyone who would favourite or retweet the tweet. In terms of the domain, question and answer platforms are used by users looking for help with a particular situation[5], however broadcasting platforms such as ours are used as ways of displaying ones thoughts, current events or generally making your voice heard. The paper "Supporting Human Decision-Making Online Using Information-Trustworthiness Metrics"[7] reflects information trust and quality metrics. They propose some new metrics that are worthy of consideration. This acted as a good read to think about different metrics for trust on social media. The Article "Size Matters" by Joshua Blumenstock[9] proposes word count as a metric for measuring article quality. While this stands true for articles on Wikipedia, on which they base their research, this may not necessarily make a huge difference in the results we get since a tweet is 140 characters regardless. "Using Readability Tests to Predict Helpful Product Reviews"[10] states that credibility features such as regularity with which bloggers post, timeliness of the posts, post length, spelling quality and appropriate use of capitalisation and emoticons in the text were found to improve performance. They specifically look at Amazon Reviews and TripAdvisor. "Credibility Ranking of Tweets during High Impact Events"[11] is quite closely related to our research, however they look at tweets of high impact events as opposed to any general tweet. They identify that characters, emoticons in a tweet, number of followers and length of username, pronouns and swear words as some of the more important content based features. The paper "Information Credibility on Twitter"[12] by Carlos Castillo looks at analysing microblog postings from trending topics and aims to classify them as credible or not credible. "Information Quality and Trustworthiness"[13] provides a state of the art review of the current work and considers the links between provenance, quality and trustworthiness. They provide a number of factors which we have used in our analysis and related these factors to Quality, Provenance or Trustworthiness.

2.2 Data Collection

The dataset used by Yahoo! Research consisted of 8,366 questions/answer pairs and 6665 questions. There were independent human editors who labelled all of the above data for quality. One would find that a drawback with such a method for labelling your data could be that it could be slightly biased based the ranker's knowledge / previous experience with the topic at hand. The Data Collection method was not discussed on the paper on Information-Trustworthiness Metrics however, they did use the dataset from the London 2011 riots both tweets from the public and news reports. For the study on Amazon and TripAdvisor they collected four large review datasets. In the paper about credibility of tweets[11] they use the Twitter Streaming API to search for key words and select tweets based on that. They Also used the Trends API which returned the trending current trending topics in twitter. They also used human annotators to rate the credibility of the tweets into four different categories. The paper by Castillo[12] looks at time sensitive information such as current news events and hence they used the $TwitterMonitor^1$ which monitors sharp increases in the frequency in a set of keywords in a message. For Classifying they used the $MechanicalTruk^2$ and asked evaluators to assist them with classification.

2.3 Analysis and Techniques

Here we focus on the types of features/characteristics chosen from the data to analyse as well as the techniques used. The experiment by the Yahoo! Research team, looked specifically at Intrinsic content quality(Punctuation, typos, syntactics and semantic complexity, grammaticality), User Relationships(User a answered a question by user b) and Usage statistics(number of clicks on something). In our research we are not necessarily looking at content quality hence we do not have these characteristics factored into our model. They also look at question quality and answer quality separately. The paper on Information-Trustworthiness metrics identifies timeliness as one of the main factors in trustworthiness. While this may be true in crisis situations or in relation to current affairs, this may not necessarily apply to general tweets, or tweets about past events. They also state that information completeness, complexity and relevance are important factors in assessing trustworthiness. They've also identified a fair few factors such as authority/reputation affecting trustworthiness, which we have factored into our model. For the experimentation they used the following factors as basis : Corroboration(extent to which the same information originates from different sources), Social-Jargon(e.g Emoticons, Shouting), Competence of the Source(Level of expertise of the source of information), Location of the source. In the paper "Information Credibility on Twitter" they separate the tweets into four types, and have Message-based features, User-based Features, Topic-based features and Propagation-based features. This is only slightly different to ours, that we don't have topic based features. They tried a couple of learning techniques such as the SVM, decision trees, decision rules and Bayes networks but achieved best results using a J48 decision tree method.

2.4 Results, Conclusions and Unfinished Work

In the paper by Yahoo! research, they selected the 20 most significant features for question quality using a chi-squared test and discovered that the precision and recall improved slightly if they used the top 10 features as opposed to all the features. The precision, recall and Area under curve for finding high quality answers was nearly

90% in comparison to the 70% obtained for question quality. This probably could be due to the fact that they used the the text as a baseline and the answer text would have more content to it than the question hence a more accurate prediction in the latter case. The article about how word count affects quality, achieved above 95% accuracy on a multi-layer perceptron, k-nearest neighbour classifier, a logit model and a random forest classifier. The paper on Credibility ranking for High Impact Events states that they found that on average 30% content about an event provides situational awareness information and 14% was spam. They also found that only 17% of the situational awareness information was credible[11]. They mention that one of the drawbacks and limitations with their data collection method was that they had to allow human annotators to annotate the classifications and would like a more automated system to establish ground truth. Our algorithm, tackles this problem by learning from known good and bad tweets.

Foot Note : 1 http://twittermonitor.net, 2 http://www.mturk.com

Part II

Model Theory

Chapter 3

Logistic Regression

3.1 Background

In the early 19th Century[1], Probit Analysis was widely used in academia and in the Industry. As time went by, the Probit Model which is closely related to the Logit Model became less popular and the latter gained popularity. The Logit Model forms the basis of Logistic Regression. Logistic Regression is a widely used and important Linear Machine Learning model. By Linear we mean we take inputs and compute a signal that is a linear combination of the inputs with weights. Although poorly named Logistic Regression, it is essentially a probabilistic classification algorithm and it builds the foundation for more complex methods like Neural networks. Logistic Regression can be used for binary classification and can be extended to multi class classification, however we will specifically look at the algorithm for binary classification since our problem is binary. In the next few pages we will look at how our algorithm works and the underlying logic behind it.

3.2 Derivation

Problem : Trust Worthiness of tweets posted on Twitter.
Structure : Our data is a collection of x and y data points where x_i is a d dimensional vector and y is a True or False classification

$$D = ((x_1, y_1), .., (x_n, y_n)), x_i \in \mathbb{R}^d \text{ and } y_i \in 0, 1$$

Input : x_1 = FrequencyOfTweets, x_2 = FavouriteCount, x_3 = RetweetCount, x_4 = InReplyToStatusID, x_5 = InReplyToUserID, x_6 = Hashtags, x_7 = UserMentionID, x_8 = URL, x_9 = MediaURL, x_{10} = MediaType, x_{11} = PossiblySensitive, x_{12} = Language, x_{13} = TweetLength , x_{14} = Location, x_{15} = Description x_{16} = UserAccountURL, x_{17} = Protected, x_{18} = FollowersCount, x_{19} = FriendsCount, x_{20} = ListedCount, x_{21} =

FavouritesCount, x_{22} = Verified, x_{23} = GeoEnabled, x_{24} = StatusesCount, x_{25} = ProfileBackgroundImageURL, x_{26} = ProfileUseBackgroundImage, x_{27} = DefaultProfile, x_{28} = IsItARetweet

Model : Logistic Regression is a discriminative model and we want to model the probability of the information being trustworthy, given some data about the tweets. We write y_i as a Bernouli Random variable with probability $\sigma(w^T x_i)$ where each of the y_i's are independent. The w here is the parameter and the x_i are fixed and non random. We give weights to each of the variables to get a linear combination.

$w_0 + w_1 x_1 + w_2 x_2 + ... + w_i x_i = w^T X$ where $X = (1, x_1, x_2, x_3)$

Here the coefficient W tells us something about how the individual variables are affecting the probability. For instance if W is negative then the probability decreases when the variable increases in value and vice versa. The coefficients also tell us which variables are more influential than the others. A large coefficient regardless of whether it's positive or negative influences the decision more thanks a small coefficient.

Let us say that the probability of the the tweet being trustworthy is P. Then the probability of the tweet not being trustworthy is $1 - P$.

The odds ratio then is : P / 1 - P We now take the natural logarithm of this : ln (P / 1-P) ln (P / 1 - P) = $w^T x$ Now if we take the exponent on both sides we get P / 1 - P = $e^{w^T x}$ P = (1 - P)*$e^{w^T x}$ P = $e^{w^T x}$ - P*$e^{w^T x}$ P + P*$e^{w^T x}$ = $e^{w^T x}$ P(1 + $e^{w^T x}$) = $e^{w^T x}$ P = $e^{w^T x}$ / (1 + $e^{w^T x}$) In this equation on the R.H.S if $e^{w^T x}$ tends to +infinity numerator in the formula below will be very huge and so will the denominator and the ratio will be close to 1. If $e^{w^T x}$ tends to -infinity both numerator and denominator would be very small and be close to 0. If $e^{w^T x}$ is 0 then $\sigma(e^{w^T x}) = 1/2$. Multiplying the R.H.S by $e^{-w^T x}$ we get P = 1 / 1 + $e^{w^T x}$

(Draw a sigmoid function in 28).

0 <= $\sigma(e^{w^T x})$ <= 1 Logistic Function : $\sigma(w^T x) = 1/1 + e - w^T x$

The Target Function f : R^d -¿ [0,1] $P(y|x) = \{f(x)fory = +1, 1 - f(x)fory = -1\}$

Our final hypothesis which we call h(x) is what we need to learn, which has the form of logistic regression, and we would claim that this is approximately equal to f(x). So we basically get the weights, multiply it with x and pass it through the non-linearity and make sure that it reflects f(x) as closely as possible. Our aim is to make this as close as possible, and what we can change here is our parameter which is W. So in other words, we know f(x), we chose Ws to give us h(x) which closely resembles f(x).

$h(x) = \sigma(wTx) = approxf(x)$

MaximumLikelihoodEstimate : Given below is how we compute the Maximum likelihood estimation for the Logistic Regression Model, for our binary scenario.
W = Parameter
D = Data
W_{MLE} = Maximum Likelihood estimate for our parameter w
$P(D|w)$ = Probability of the Data given w
$P(y|x)$ = Probability of y given x

$W_{MLE} \in argmax\ P(D|w)$

$P(D|w) = \prod_i = 1^n P(y_i|x_i, w)$
We can write the probability mass function of a bernouli as : $= \prod_i = 1^n \sigma(w^T x)^{y_i}(1 - w^T x)^(1 - y_i)$

A brief comparison of Logistic Regression with some other classification models are given below :

Linear classification:

We take our hypothesis to be a decision, and this decision (+/- 1) is a direct result of signal. In the following diagram our inputs are x1?xd where X(dash on top) is a d dimensional vector and we sum it and pass it through the following threshold to get a decision. Insert Diag 1

Linear Regression

In the case of Linear Regression we do nothing to the signal. Here we output what we input, the input being the sum of vectors x1..xd. Insert Diag 1

Logistic Regression

In Logistic regression we will take our sum and apply a non-linearity to it. Here theta the logistic function is not as harsh as the Linear Classification and the output would be interpreted as a probability. Again in our diagram we have a d dimensional vector x and the sum is passed through this function to give us a probability. insert Diag 1

Pros :

1.Small number of parameters : In the above the number of parameters would be (d+1). What this means is that even if the dimensions increase, the number of parameters only linearly increases. 2.This model is Computationally efficient to estimate the w parameter. We use Newtons method which is a very efficient second order method which is guaranteed to converge to a Maximum Likelihood estimate. 3. You can extend it to multi class classification 4. Forms the foundation for some of the more complex methods like neural networks.

Cons : Performance is not necessarily as good as some of the best performing methods such as random forests, boosted methods and support vector machines. This however, greatly depends on the nature of the problem and from what we can see this has performed quite well for our problem.

Chapter 4

Support Vector Machines

Part III

Experimental Setting

Chapter 5

Method

Each tweet on twitter has two distinct components : The Actual Tweet (Includes the content of the tweet and the meta data associated with it) and the User(Characteristics of the user).

Part IV

Appendix