Initial Plan

Football Transfer Rumours on Twitter: Who and What Should We Believe?

Author: Wiliam Thomas - C1519266

Supervisor: Richard Booth

Module Number: CM3203

Module Title: One Semester Individual Project

Credits: 40

Submission Date: 05/02/2018

Project Description

2017 is a year that will go down in history for its association with the rise of fake news. 2017 saw the term fake news being tossed around by everyone directed at anyone, it's ubiquitous presence caused a stir in the international world. It's use proved to be hugely damaging to the entire journalism industry, from freelance journalist to multinational news outlets present worldwide. The origin of the word 'fake news' is a cause for controversy, as is its outbreak to the public eye in the year 2017. There is little doubt that 2017 was the year that truly caused the accusations of 'fake news' to be taken seriously and for such claims to be made at the frequency it did, up to the point that 'FakeNews' was coined as 'Word of the Year' by Collins Dictionary's lexicographers. (1)

The issue of fake news and what/who to believe is a problem as old as civilisation itself. It is not the case that suddenly reporters have suddenly lost all honesty and integrity in the space of a year. Perhaps honesty & integrity is not the issue in these cases but rather interpretation and how news are worded, or perhaps the issue lies with the audience of news reports and their awareness of Fake News has suddenly caused an uproar of trust issues. Whatever the cause is, there is no doubt a crisis of trust in the industry. The issue of who to believe and trust might have only come to light in the recent year for the average person, but this is an old and longstanding problem in the world of Football Transfers. For anyone who follows football, fake news is as real as the game itself, football journalism has been plagued by fake transfer rumours for years and the problem of which news source to believe and not believe is anybody's guess. Football Transfers is often called the original fake news, and with good reason. Small freelance journalist have traditionally struggled to gain a following, but the invention of social media has changed all that. It's ability to publish articles to potentially millions of followers has made it easy for freelance journalist to gain a following but as a result it has made it easier for fake news to flourish. (2)

Football Transfers worldwide is a huge area, this project limits the scope to focus only on transfers occurring within the English Premier League, extending the scope beyond this limitation would be a logical next step in the project given more time and resources, but for now, this project aims to be a workable proof of concept rather than a solid 'go-to' tool for dealing with beliefs and reliability. With Twitter being the go to social media platform for journalists publishing transfer rumours and the abundant availability of twitter data to the public, this project aims to quell fears of fake news in the football transfers industry by processing large amount of twitter data using python modules and Twitter's python API and utilising Natural Language Processing Techniques demonstrated in the book Natural Language Processing with Python. (3) The result of processing thousands of individual tweets will be records of formal logic statements that takes the form of "User U claims player P will join team T". The project then aims to implement existing algorithms to formulate a reliability score for each source in the dataset, which will be computed based on whether their statements became reality alongside a belief value on any given statement of the form "player P will join team T" at any given time. Part of the program is to research suitable existing algorithm to solve this problem. Given the scale and current popularity of

the fake news problem, it is not a unique problem as demonstrated in the paper Knowing What To Believe by Pasternack (4), but solving it in the context that I have proposed will require manipulating existing algorithms to work for this particular problem. Alongside existing algorithms, I hope to investigate a unique and hopefully better solution and propose improvement on the existing algorithms to come up with a potentially better solution. The system itself will be implemented as a command line interface system, where it will be possible to import a dataset and a user will be able to see the most reliable sources and see the likelihood of a player moving from team A to team B at any given time set by the user. Alongside the implemented system will be a written report documenting the steps taken to get the final output along with the recommended algorithm and improvement and how the system should be utilised to be used concurrently in future transfer periods.

Aims and Objectives

Aim 1

Process tweet dataset of JSON Documents into individual records containing date/time, username, user_ID, tweet_text. Everything else from the JSON document is to be filtered out. These records should then be further processed into a final .csv file. This file output should contain records sorted in date/time order where the oldest tweets appear first.

Objectives

- Study API Document of Tweepy Module
- Become familiar with the JSON Document format
- Create Python script that takes original dataset as input
- Implement algorithm to filter out attributes not required in every JSON Document
- Output every JSON file as a record containing the attributes : User_ID,User_Name,Tweet_Text
- Generate a CSV file, with each row containing a record generated from the previous objective

Aim 2

Using Natural Language Processing Techniques and the NLTK module in Python, extract relevant tweets from the dataset to create formal logic statements of the form : “User **U** claims player **P** will join team **T**”

Objectives

- Study general Natural Language Processing Methods
- Study API Document of NLTK Module
- Read the book Natural Language Processing with Python (3)
- Create Python script that takes CSV of processed tweets as input

- Implement Natural Language Processing Methods using the NLTK module on python to process each twitter text attribute of every record to break down text into 2 parts (Player **P**, Team **T**)
- Output statements in suitable format (Format to be decided via further research)

Aim 3

Research suitable existing algorithms to implement a method which, at any given time computes and outputs a **Reliability** value for every twitter user in the dataset based on their past predictive statement(s)

Objectives

- Background research on similar projects
- Study methods proposed in the paper on "Knowing What To Believe" (4)
- Create Command Line Python program which utilizes the script implemented for Aim #2 and takes the output generated from that script as the input for a method to compute a reliability score
- Modify an existing algorithm to work with the input data and return a reliability value as a decimal (Between 0 and 1) for every user_ID present in the input data

Aim 4

Research suitable existing algorithms to implement a method which, at any given time computes and outputs a **belief** value for any predictive statement of the form "Player **P** will join team **T**" based on the **Reliability** values of Users **U** who have/haven't made a statement on the Player **P** joining team **T**

Objectives

- Background research on similar projects
- Modify an existing algorithm to work with the input data and return a Belief value as a decimal (Between 0 and 1) for the selected statement (Of the form Player **P** will join Team **T**)

Aim 5

Create Python Command Line program which uses existing algorithms and from these algorithms is able to output a reliability value for every twitter user in the list of statements (.csv file) along with a belief value for any given statement of the form Player **P** will join Team **T** given that the player is registered in the English Premier League & the team is a competing team in the English Premier League

Objectives

- Implement both algorithms developed for Aims 4 and 5 in Python
- Implement a command line user interface with the following options
    - Sort Twitter Users by reliability
    - Search for Twitter User and their Reliability & History of statements
    - Calculate what was the belief value of a Player **P** moving to Team **T** at a given date & time

Aim 6

Research and propose a new original algorithm/method to determine potentially more accurate reliability and belief values

Objectives

- Research different possibilities and different methods of computing reliability and belief values
- Come up with different formulas for computing reliability and belief values and test these different formulas to obtain experimental results
- Compare experimental results with original result outputted using existing methods
- Evaluate which algorithms perform better under the circumstances

Work Plan

The workplan has been developed with 15 weeks in mind, starting from the **29/02/2018** up to the submission deadline on the **11/05/2018**.

Week	Task	Milestone
Week 1 29/02/ – 04/02	➢ Discuss Ethical Approval arrangements ➢ Complete online Research Integrity Module ➢ Complete Initial Plan	➢ Awarded a Certificate of Completion of the Research Integrity Module ➢ Initial Plan Submitted
Week 2 05/02/ – 11/02	➢ Background Research & Reading in NLP ➢ Start reading of Natural Language Processing with Python book	➢ General understanding of the NLTK (Natural Language ToolKit for Python) ➢ Ability to program basic NLTK commands in Python
Week 3 12/02 – 18/02	➢ Create a Python Script to process JSON Dataset ➢ Process JSON Documents to the correct form using Python Script created in order to achieve Aim 1 ➢ Continue Reading of Natural Language Processing with Python	➢ .py file which contains code to process dataset into desired output A .csv file which contains relevant JSON Documents from the dataset processed to be in the form of a record which contains the users ID, the username and the tweet **AIM ACHIEVED** : Aim 1
Week 4 19/02 – 25/05	➢ Finish researching into Natural Language Processing ➢ Write a summary on Natural Language Processing and how to apply what I have learn to the project ➢ Finish NLTK Python Book	➢ Solid understanding of NLP and the NLTK for Python ➢ A section on NLP and its applications written for the final report
Week 5 26/02 – 04/03	➢ Implement NLTK Python script to convert records into formal logic statement of the form Player **P** will join Team **T** ➢ **Review Meeting 1 with Supervisor**	➢ Output (Format yet to be decided) where relevant tweets have been extracted and converted to the format: User **U** claims Player **P** will join Team **T** **AIM ACHIEVED** : Aim 2
Week 6 05/03/ – 11/03	➢ Background Research into suitable algorithms to use ➢ Study paper Knowing What To Believe in depth	➢ Greater understanding of existing solutions to similar problems
Week 7 12/03 – 18/03	➢ Implement algorithm to calculate reliability value of all twitter users in the dataset	➢ Working python script that calculates reliability values for all twitter users in dataset **AIM ACHIEVED** : Aim 3
Week 8 19/03 – 25/03	➢ Implement algorithm that calculates any given belief value at any given time	➢ Working python script that calculates belief values for any statement at any given time **AIM ACHIEVED** : Aim 4
Easter Recess 26/03 – 15/04	➢ Combine Implementation of both algorithms to build simple command line tool that allows searching / sorting of different users and/or beliefs	➢ A working program in Python that allows users to import dataset and the program outputs a list of users and their reliability score & prediction history. Program should also allow users to search for a user to find their reliability score **AIM ACHIEVED** : Aim 5
Week 9 16/04 – 22/04	➢ Research and test new original ways of calculating belief and reliable values ➢ **Review Meeting 2 with Supervisor**	➢ Variety of different possible solution outlined and tested against the dataset
Week 10 23/04 - 29/04	➢ Compare current method and new method of calculating belief and reliable values and come to conclusion	➢ Experimental results along with written evaluation of different algorithms and how well they work **AIM ACHIEVED** : Aim 6
Week 11 30/04 – 06/05	➢ Compile different areas of report together into correct format ready for submission	➢ Final Report will have been finished and proof read ready for submission
Week 12 - Deadline 07/05 – 11/05	➢ Final week to fall back on if any problems arise	

References

1. **Flood, Alison.** *Fake news is 'very real' word of the year for 2017.* s.l. : Guardian, 2017.

2. **Smith, Rory.** *The Original Fake News: Soccer Transfers.* s.l. : New York Times, 2017.

3. **Steven Bird, Ewan Klein, and Edward Loper.** *Natural Language Processing With Python.* s.l. : O'Reilly Media, 2009.

4. *Knowing What to Believe (when you already know something).* **Jeff Pasternack, Dan Roth.** 2010.