TriVis: Visualising Multivariate Data from Sentiment Analysis

Maryanne Doyle*
University College Dublin
maryanne.ni-dhuaill@ucdconnect.ie

Alan F. Smeaton
Insight Centre for Data Analytics
Dublin City University
alan.smeaton@dcu.ie

Adam Bermingham
Insight Centre for Data Analytics
Dublin City University
adam.bermingham@dcu.ie

ABSTRACT

In a time when a single sporting event can elicit millions of Tweets the volume of expressions of sentiment available is far too large to be read by an individual in real time. TriVis is a visualisation design that uses a modified scatter plot with three axes to allow the user to read and understand multidimensional data at a glance. We examined the readability of the visualisation using data collected from a golf tournament and plotted the sentiment towards golfers in real time during play. TriVis visualisations are simple, easy to understand and offer insights into the data set which are not obvious using other methods.

Categories and Subject Descriptors

H.5 [**Information Interfaces and Presentation**]; H.5.2 [**User Interfaces**]

General Terms

Data visualisation, design

Keywords

Visualisation, sentiment analysis

1. INTRODUCTION

Consider a spectator watching a golf tournament who witnesses a favourite getting knocked out of the competition. If this user wishes to understand how other spectators reacted to these events they might consider a gasp from the attendees, a criticism from the commentators or await a post-game analysis by experts. In the age of social media we have access to a rich data set of the reactions of millions of people who micro-blog on platforms such as Facebook, Tumblr and Twitter. This data set requires a computational approach such as text analysis before the results can be represented using a visualisation to convey the information to our user. The challenge is to present this analysis in real time in a form that is immediately understandable.

Using Twitter a user can rapidly read and author content without significantly interrupting their viewing experience resulting in large volumes of related Tweets during sporting competitions [1]. This compatibility with sports commentary makes Twitter an ideal platform from which to harvest data for our test case. This data is easily collected but we need to ask ourselves which elements are of interest to spectators, players, sponsors or the general public. When was the Tweet authored? How many people Tweeted? Were the authors' reactions good or bad? These reactions can be classified using sentiment analysis to determine whether a given Tweet was positive, neutral or negative. Sentiment is expressed when a user authors content that is subjective, opinionated or emotional and includes evaluation or speculation [1].

For large data sets visualisation is often a more effective way of describing the data than a table or a purely numeric representation [2]. Well designed visualisations are easier to read than tables or spreadsheets, especially for non-technical users, and can expose patterns in the data that would be difficult to see otherwise. An important principle in visualisation design is the data-ink ratio [2]: ideally a visualisation will convey a large amount of data through the use of a small amount of ink or design elements. Design elements such as shape, colour, size and position can be assessed by the user in under 250 milliseconds using preattentive processing [3]. Designing a visualisation using these elements and avoiding clutter to achieve a good data-ink ratio will allow our user to read the data at a glance.

Although visualisations such as bar charts and line charts are relatively easy to read, they struggle to convey multidimensional data since the addition of multiple y-axes complicates the display and compromises readability. Similarly separating the data by using multiple line charts gives the user more work to do to piece together the full picture. Visualisations such as scatterplots or networks are more suited for the display of multidimensional data but require an investment of time and some examination by the user to extract meaning. TriVis uses a modified scatterplot with three axes to show the full range of sentiment classifications in the data set.

In this section we have presented the difficulties in visu-

*Work carried out while at Insight Centre for Data Analytics in DCU

The 8th Irish Human Computer Interaction Conference Dublin City University, Glasnevin, Dublin 9

[1]Wimbledon causes surge in social media. `http://www.bbc.com/news/technology-23225501`

alising high dimensional data and the motivation to design a solution to gain understanding from large data sets. In the next section we will discuss related work and examine progress in the area of visualisation of sentiment with respect to the design of TriVis. In Section 3 we provide a walk through of plotting data using TriVis before a discussion in Section 4 using example visualisations. Finally we outline future work in Section 5 and draw conclusions in Section 6.

2. RELATED WORK

Some work on sentiment analysis uses a large number of sentiment classifications such as that by Fukuhara et al. in temporal analysis of social events [4]. In this work a stacked line diagram is used to show the classifications anxiety, sorrow, shock, complaint, anger, happy, fatigue and suffering. The focus is on exploring events and sentiment over time but the visualisations are limited to showing either a range of sentiment classifications for a single event or a single sentiment classification for a series of events.

Visualisations such as TwitInfo [5] and Vox Civitas [6] use two separate graphs to show volume and sentiment while others such as VISA [7] combine the two into one graph. Including volume is preferable since the user can see at a glance if the sentiment score represents the views of many authors who are in agreement or few individuals who are outliers in the data set. The visualisation used in VISA makes it somewhat difficult to compare volume from one topic to another, although the overlaying of keywords is a useful addition. TwitInfo uses only two classifications of sentiment (positive and negative) while Vox Civitas adds two more (neutral and controversial) although both have added an element of event detection that allows for further exploration. More recent work such as Semantize [8] conveys sentiment found in a document with the use of font and background colours which are used to highlight the text.

TriVis uses three separate classifications of sentiment (positive, neutral and negative) to represent multidimensional social sentiment data. This design combines sentiment and volume in the same graph can show multiple topics distinguished by the use of colour. TriVis shows a distinction between sentiment labelled as neutral and sentiment which is composed in equal amounts of negative and positive. The latter represents a controversial or polarised sentiment and is graphed in a separate region of the visualisation to neutral sentiment. TriVis does not use time on the x-axis but instead uses a non-linear representation where time can be seen as a series of pathways which join plotted sentiment values.

3. PLOTTING SENTIMENT

In this section we outline how data was collected for the examples used in this paper, describe the algorithm for plotting sentiment and discuss how the data is represented visually.

3.1 Data Collection

The sporting event used to collect test data and test the system was the Open, a golfing tournament held in Scotland in July 2013 [2]. We collected Tweets in real time using the Twitter Streaming API and filtering for the names of golfers who took part [3]. Next these Tweets were analysed using a supervised machine learning algorithm [9] which attributed a label to each Tweet indicating whether the sentiment expressed towards the topic or golfer was positive, neutral or negative. Sentiment towards a golfer for a given time period is the relative frequency of positive, neutral and negative labels for that time. The goal is to design a visualisation that presents this analysis to a spectator of a live event such as the Open in an easily readable visualisation. This visualisation could then be overlayed on live play or added to a leaderboard to show sentiment towards the players on screen and in the tournament as a whole respectively.

3.2 Algorithm for Plotting Sentiment

After the steps outlined in 3.1 we produced sentiment tuples.

$$Sent(T)_i =< pos, neu, neg > \quad (1)$$

where pos, neu and neg are the number of Tweets that were positive, neutral or negative that exist for a given golfer or topic T at time interval i. The first step was to normalise these values as shown in Equation 2 were pos is normalised. The sum of pos, neu and neg is always equal to 1.

$$posnorm = \frac{pos}{pos + neu + neg} \quad (2)$$

These normalised values were converted to a plotable form using polar coordinates (r, θ) mapping to a conceptual wedge shape (of θ 90 degrees). The normalised neu value was used to map to r directly, meaning the height of r corresponds to the neutral value, the length of r increasing as neu decreases. θ was used to convey both pos and neg using Equation 3.

$$\theta = 90 * (\frac{neg}{neg + pos}) \quad (3)$$

This produced an angle of 0 which corresponded to the left most side of the wedge for a 100% negative score and an angle of 90 corresponding to the rightmost side for a 100% positive score.

Multiple coordinates with the same value for r would appear at different heights as the value for θ changed. We scaled the value of r to flatten out this curved presentation using Equation 4.

$$r_{scaled} = r * cos(\theta) \quad (4)$$

Data Component	Visual Dimension
positive sentiment	x,y coordinates
neutral sentiment	x,y coordinates
negative sentiment	x,y coordinates
time	trail
topic	colour
volume	area

Table 1: Mapping of data to display dimension.

[2]Muirfield - 2013 Results `http://www.theopen.com/en/History/OpenVenues/Muirfield.aspx`

[3]Twitter Streaming API `http://https://dev.twitter.com/docs/api/streaming`

3.3 Visual Representation

Once plotted in polar coordinate form the sentiment tuples appear as points within a triangular area. Each of the vertices represents a sentiment classification and proximity to a vertex indicates the composition of sentiment of each tuple. Points appearing close to the positive vertex have a higher positive score than negative or neutral. Points midway between the positive and negative vertices have a low neutral component and represent polarised topics where the positive and negative scores are of close to equal value. Multiple topics are shown using a different colour and the volume of Tweets associated with any plotted point is shown through the area of the point itself. Trails connecting points indicate a path over time. Table 1 lists how the data components map to the visual dimensions used.

The 2D visualisations shown in this report were produced using a JavaScript graphing API Highcharts [4] and the 3D visualisation was produced using three.js, a library for WebGL [5].

4. RESULTS AND DISCUSSION

In this section we discuss how sentiment data is displayed using TriVis and use a series of examples from our sample data set to explore the readability of the graph.

4.1 Sentiment in 2D

Figure 1 shows the basic template of the visualisation, a 2D triangle where each vertex represents a maximum score in one of the three SentiSense classifications of sentiment: positive, neutral or negative. A topic whose sentiment score was split evenly over all three classifications would appear in the centre of the triangle, while a score dominated by one classification is plotted closer to that corner than either of the other two. The further from the neutral vertex a point is plotted the more polarised the sentiment on that topic.

4.2 Examples

In Figure 2 the sentiment score representing Rory McIlroy (shown in orange, ranked second in the world at the time the Open began [6]) veers towards negative as he plays poorly on days one and two and fails to make the cut to progress to the weekend. Sentiment towards Phil Mickelson (shown in blue, ranked fifth in the world at the time) hovers on the positive side of neutral before surging towards positive and growing dramatically in volume after he wins the tournament on day four. In Figure 3 we see that sentiment towards Tiger Woods (shown in purple) was composed predominantly of positive and negative values which indicates that opinions of his performance and prospects in the tournament were either highly positive or negative. If plotted on a scale of negative to positive this type of score would appear in the middle and give the impression of neutrality rather than polarisation which is the case. Using TriVis polarised sentiment is clearly distinguishable from neutral sentiment. The volume of Tweets about Tiger Woods is substantial and is larger than that of Ian Poulter shown in green until the final day when he rose to third place and became the highest ranking Briton in the final hours of the game.

[4]Highsoft AS `http://www.highcharts.com/`
[5]Three.js Library for WebGL `http://threejs.org/`
[6]Official World Golf Ranking `http://www.owgr.com/Ranking`

4.3 Representing time

As the representation of time is non-linear temporal patterns appear as a footprint of pathways and points over time. Saturation of colour is used to indicate the passage of time as shown in Figure 4 where older values are more faint. Another representation is to align a series of snapshots of single time points along the z-axis, using a third spatial dimension to convey time as shown in Figure 5 although this representation is best explored by a user who can rotate and examine the 3D model.

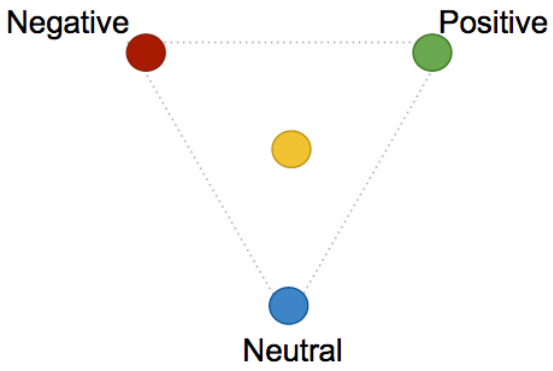

Figure 1: 2D template for plotting sentiment.

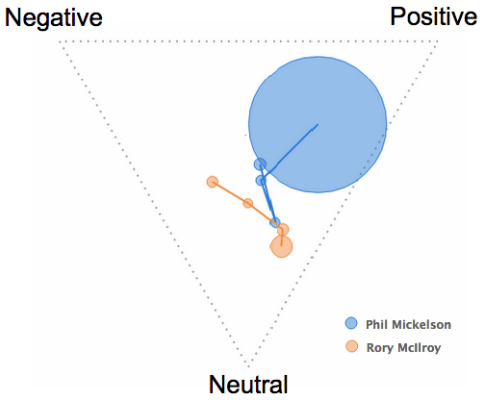

Figure 2: Sentiment towards Phil Mickelson and Rory McIlroy with daily intervals.

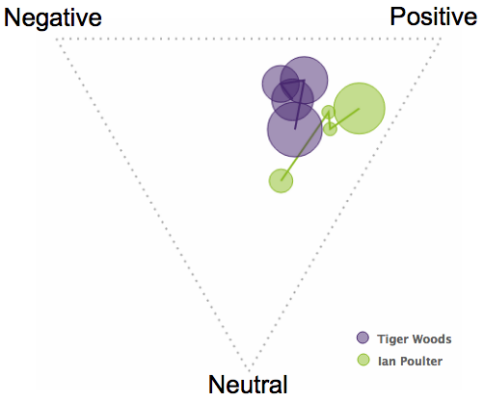

Figure 3: Sentiment towards Tiger Woods and Ian Poulter with daily intervals.

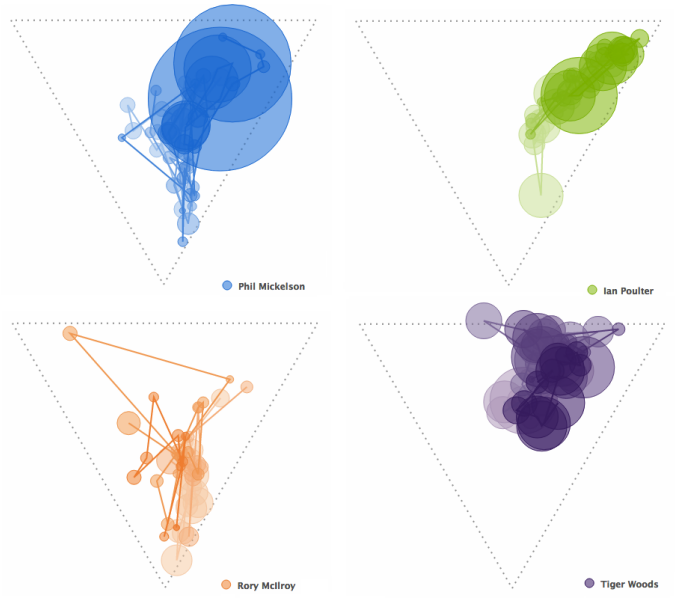

Figure 4: Sentiment towards Phil Mickelson, Ian Poulter, Rory McIlroy and Tiger Woods woth hourly intervals.

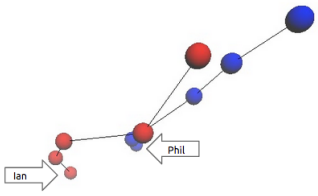

Figure 5: Daily values for Phil Mickelson and Ian Poulter plotted in 3D (sphere size does not represent volume).

In Figure 4 the transparency of the circles representing volume allows the user to see detail beneath the most recent layer, but also creates some interference as layers of paths and volume circles appear darker and more recent as a result. These footprints are useful in seeing a general profile or sentiment pattern in a large data set.

5. FUTURE WORK

We would like to investigate the application of this approach in other areas such as political debates, news, talent competitions etc. Also of interest is exploring the use of this design as an addition to a leaderboard or scoreboard, as a symbol to overlay on live television and the use of animation to update the visualisation in real time. A natural extension of this work would be to implement interactivity as part of a drill down interface to allow the user to have more control over the visualisation by zooming or filtering to explore specific times or keywords. This would allow the user to see an overview using TriVis in its current form, then zoom and filter to explore details on demand covering all aspects of the visual information seeking mantra [10].

6. CONCLUSION

In this paper we explored the challenge of representing high dimensional data using an easily readable visualisation. The data set we used was captured from Twitter during a golf tournament and analysed for positive, neutral and negative sentiment. We examined other work carried out in visualising sentiment data and designed and implemented a novel visualisation of our own. TriVis represents volume and composition of sentiment towards multiple topics over a given interval of time in one graph. TriVis can be read at a glance by a spectator of a live event and provides an understandable insight into a complex data set in real time. The implementation does not facilitate interactivity at this time but could be extended to allow a user to filter and explore the data set in more depth.

7. ACKNOWLEDGEMENTS

This research was supported by Science Foundation Ireland under grant number SFI/12/RC/2289 and Enterprise Ireland under grant number CF/2012/2618.

8. REFERENCES

[1] B Pang and L Lee *"Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval"* 2(1-2):1135 2008

[2] ER Tufte. *"The Visual Display of Quantative Information 2nd Edition"* 2001 ISBN 0-9613921-4-2

[3] C G Healey, K S Booth, J T Enns. *"Visualizing real-time multivariate data using preattentive processing"* 2001 ISBN 0-9613921-4-2

[4] T Fukuhara, H Nakagawa, and T Nishida. *"Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events."* ICWSM. 2007.

[5] A Marcus, M S Bernstein, O Badar, D R Karger, S Madden and R C Miller. *"Twitinfo: aggregating and visualizing microblogs for event exploration."* In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 227-236. ACM, 2011.

[6] N Diakopoulos, M Naaman, and F Kivran-Swaine. *"Diamonds in the rough: Social media visual analytics for journalistic inquiry."* In Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on, pp. 115-122. IEEE, 2010.

[7] D Duan, W Qian, S Pan, L Shi, and C Lin. *"VISA: a visual sentiment analysis system."* In Proceedings of the 5th International Symposium on Visual Information Communication and Interaction, pp. 22-28. ACM, 2012.

[8] AJ Wecker, J Lanir, O Mokryn, E Minkov, T Kuflik . *"Semantize: Visualizing the Sentiment of Individual Document."* AVI '14 Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces pp. 385-386 2014.

[9] A Bermingham and A F Smeaton. *"Classifying sentiment in microblogs: is brevity an advantage?."* In Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1833-1836. ACM, 2010.

[10] B Shneidermen *" The eyes have it: a task by data type taxonomy for information visualizations."* IEEE Symposium on Visual Languages p336 - 343.