Text Analysis: Sport News

Introduction

Sport journalism is playing an important role in public awareness regarding national pride, athletic performance and the socio-political in sports. This is crucial because it can influence notional performance and pride. When athletes perform poorly it can bring shame to the country. Likewise, management mistreats athletes of fails to support them properly it may lead to national embarrassment and discourage new talent from contributing to country's sporting success.

This project focuses on text analysis of 120 Malaysia sports news articles and written in English and sourced from reputable online platforms which is New Straits Times (NST) Malaysia and The Start (Malaysia). This news covering a wide range of sports in Malaysia which are football, badminton, hockey, cycling, tennis, bowling, esport and more. The objective is to explore underlying themes, sentiments and language patterns inside the sport news using text mining techniques. Several analysis like topic modelling, clustering, sentiment analysis using Bing and NRC lexicons and TF-IDF analysis will be done to uncover dominant topic across articles and its insight such as emotional tones convey in the news and others. This is needed to understand how Malaysia sports achievements, controversies and the way athlete communicate with public like in international tournaments.

Methodology

In this project, data for 120 sport news was collected manually from the online platform. These news consists of various sports domain such as football, badminton, hockey, cycling, golf, esport, bowling, tennis, athletics and others. Then, text data was cleaned and normalized using standard preprocessing which are conversion to lowercase, removal of punctuation and numbers, elimination of English stop words and stripping whitespace. Furthermore, lemmatization was applied instead of stemming because to preserve meaning while stemming may produce incomplete or incorrect root forms.

Then, constructing Document-Term Matrix (DTM) and topic modelling using Latent Dirichlet Allocation (LDA). In this project two model was trained which are k =3 for general insight and use library = 'ldatuning' to find a optimum k value. Here are using metrics like "Griffiths2004", "CaoJuan2009", "Arun2010" and "Deveaud2014". Figure 1 shows linear modelling tuning by topic.

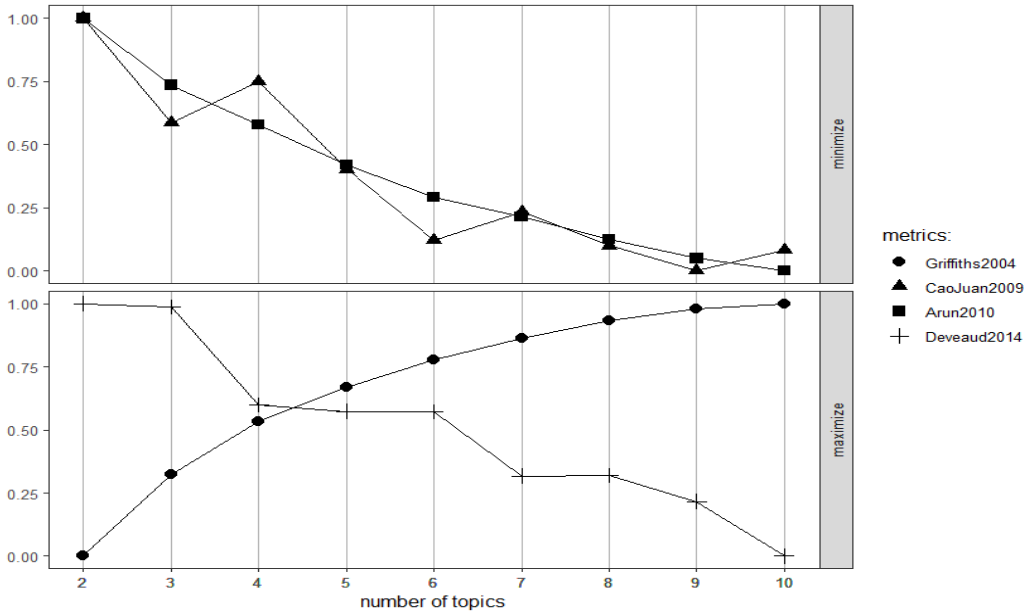

Figure 1 Linear modelling tuning

From the graph above, the optimal k value get are k =6 for more fine topic separation. Top 10 frequent words will be plot and compared. Next, sentiment analysis was done using Bing and NRC lexicons to identify positive and negative emotion. Lastly, TF-IDF is performed to identify unique keywords or phrases that distinguish each new from others.

Analysis and Discussion

There are two LDA topic modelling were conducted using k=3 and k=6 on a collection of text document. This approach groups the news into three or six main topics to allow better understanding of the dominant themes within the dataset. Each group represents a different topic such as Topic 1 is about football and Topic 3 could be about badminton. Figure 2 below shows top 10 terms per topic for k=3.

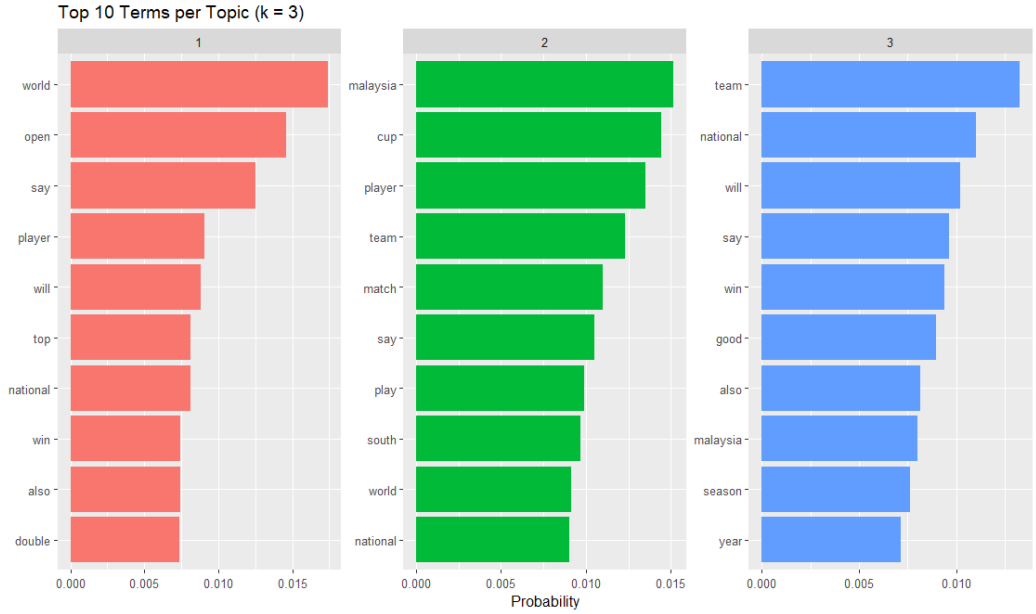

Figure 2 Top 10 terms per topic for k=3.

From Figure 1, the news was grouped into three themes. Topic 1 appears to have global context such as ‘world’, ‘open’ and ‘national’ where it can be said athlete participate in big tournaments like Piala Sudirman for badminton and usually tournaments in badminton are using ‘open’ in their competitions names. Then, Topic 2 is a team focused gameplay like football and double-man or woman badminton. This is because in Topic 2 there are words like ‘team’ which showing a group gameplay and ‘cup’ regarding to the name of the competition like in the football or badminton. It also emphasize in national theme where there are words ‘Malaysia’ and ‘national’ inside it. Next, Topic 3 is leaned towards the performance or progression of the athletes because ‘season’, ‘year’, ‘good’ and ‘win’ shows the athletes are prepared for the next tournament or them talk about their past performance. However due to the smaller number of topic, some theme might be overlap like terms ‘national’ have at all topic. Figure 3 below shows top 10 terms per topic for k=6.

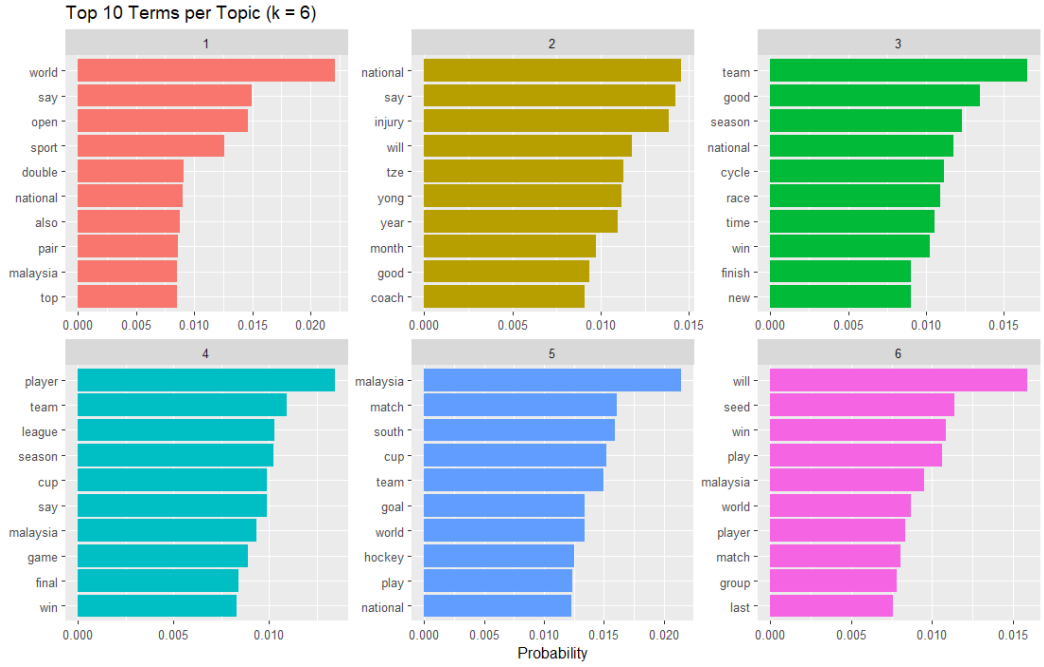

Figure 3 Top 10 terms per topic for k=6.

Based on Figure 2, when increasing to k=6, more topic are appear and more refined and easier to distinguish. For example in Topic 1, the theme is about badminton and world level performance due to terms like ‘world’, ‘open’, ‘badminton’, ‘pair’ and others. All the terms is showing the badminton performance which can be pair or double group performance. It also talk about how badminton is in top of the world. Then, Topic 2 is centred around injuries and recovery where terms like ‘injury’, ‘coach’ and ‘month’ appear. This can be Tze is injured on this month and coach predict Tze will perform back next month.

Next, Topic 3 clearly focused on cycling event where top terms like ‘cycle’, ‘race’, ‘time’, and ‘finish’ appears. This might be new about the performance or the national cycling where the data like race time and time finish written in the news. Then, Topic 4 is representing football leagues as seen from ‘league’, ‘season’, ‘cup’ and ‘player’ as top terms. Most of footballer was called like player 7 based on their shirt number during games.

Topic 5 dealt with international hocket and team tournaments which reflected by ‘goal’, ‘hockey’ and ‘match’. Hockey Malaysia one of the well-known team in international levels due to their gameplay and how pro there are. Finally, Topic 6 seemed to capture mixed content around individual matches with terms like ‘seed’, ‘last’ and ‘play’. It can be from other sports like bowling, ping pong and others. The increased number of topics allowed better separation and showing different sports issues like injuries, tournaments or type of sports that covered

widely in Malaysian media. Figure 4 show the different beta spread between Topic 1 and Topic 3 for k=3 model with threshold 0.005.

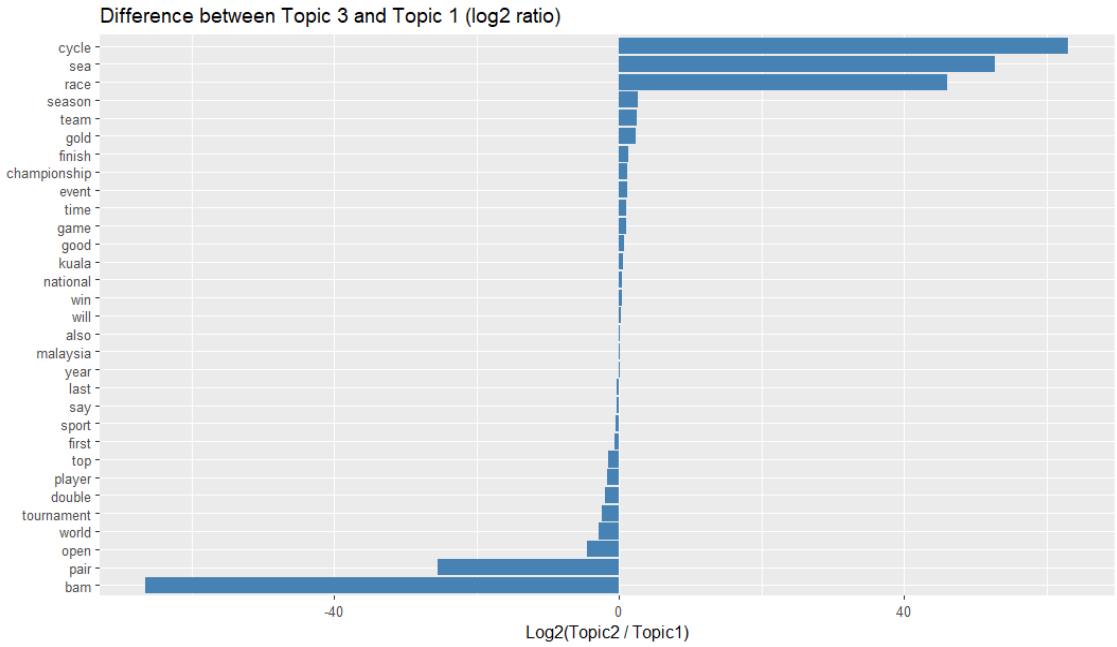

Figure 4 Difference between Topic 3 and Topic 1 for k=3.

To understand the difference between Topic 3 and Topic 1, compared per-topic probabilities (Beta values) and computed log2 of Topic 3 over Topic1. Positive log ratios represent Topic 3 while negative values represent Topic 1. Topic 3 appears to centre more ion endurance sport or racing events like cycling due to high probability terms of 'cycle', 'sea' and 'race'. Additionally, terms like 'team' and 'good' are slightly more associated with topic 3 which suggesting evaluation and group based events. On the other hand, Topic 1 is definitely talking about badminton because the highest probability term is 'bam' which stands for Badminton Association Malaysia (BAM). This supported by terms like 'open', 'tournament', 'world', 'double', 'top' and 'pair' where there are Malaysia athlete is ranking in number one and number two pair or double man in the world in recent open tournaments. Some neutral or shared terms with small log ratios include 'Malaysia', 'national' and 'win' showing both topics which indicate the news about athlete compete at outside of Malaysia for both topics. From here also can see the distinct between both topic at the highest probability terms of each topics but Topic 3 have more broad sport compared to Topic 1. Figure 5 show the different beta spread between Topic 1 and Topic 5 for k=6 model with threshold 0.005.

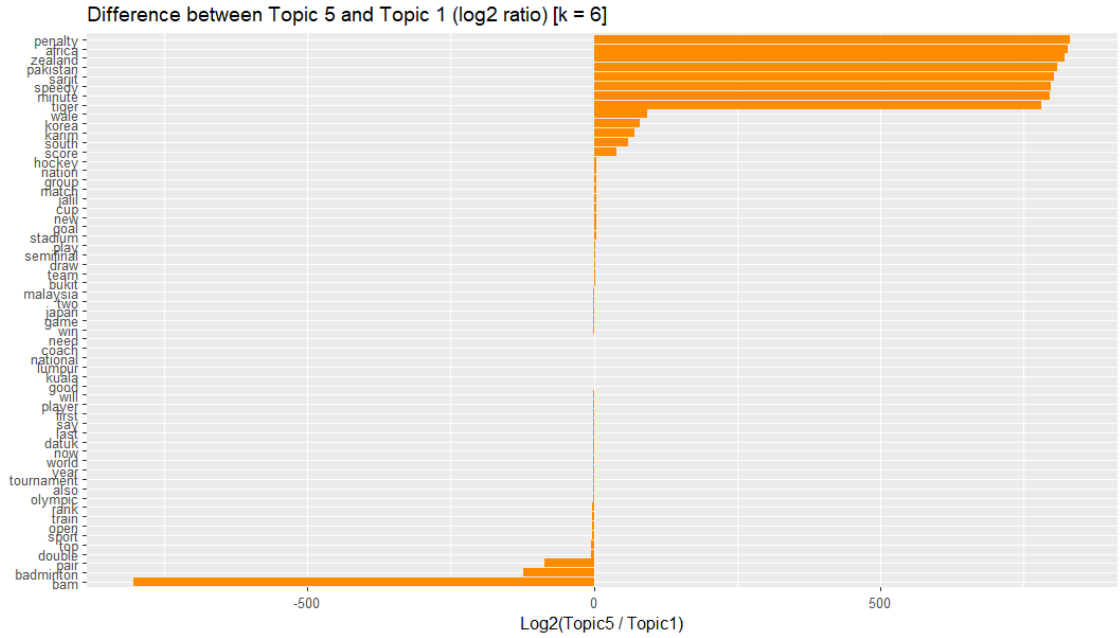

Figure 5 Difference between Topic 5 and Topic 1 for k=6.

Based on Figure 4, the topic-term log ratio comparison where negative is representing Topic 1 and positive represent Topic 5. Topic 5 appears to dominate sport and international references with highlight probabilities terms for ‘Africa’, ‘Korea’, ‘Pakistan’, ‘penalty’, ‘minutes’, ‘match’, ‘goal’ and ‘speedy’. The ‘speedy’ terms is representing Malaysia hockey team and this is show Topic 5 is news related to the world hockey competition which happening right now in Malaysia. Then, Topic 1 is talking about badminton in Malaysia due to high probability terms of ‘BAM’ and ‘badminton’. As compared to model k=3, Topic 1 here is more details where the second highest probability here is badminton at -123 which really show the main topic only about the badminton while at model k=3 does not have term ‘badminton’. Then, k=6 model is a distinct model where every topic have their own sport new like Topic 4 here is about hocket and Topic 1 is about badminton while at k=3 model, Topic 3 can be any sport because term like ‘cycle’ and ‘sea’ are contradict to each others. Figure 6 shows the sentiment analysis for both model using Bing and NRC lexicons.

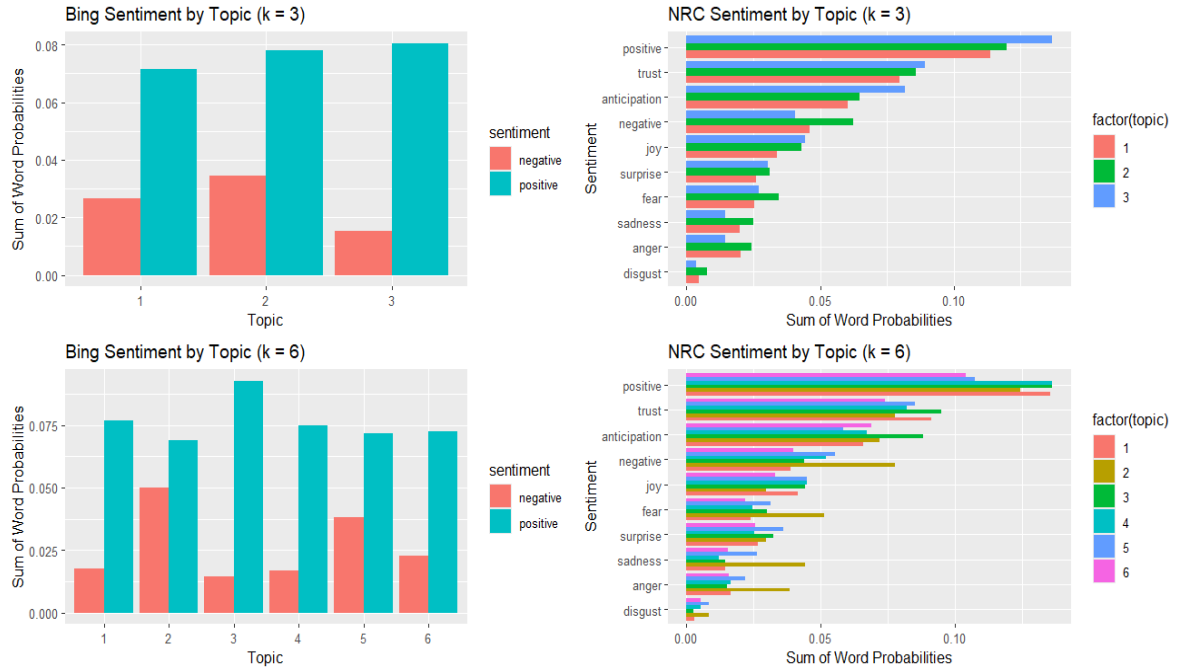

Figure 6 Sentiment analysis using Bing and NRC lexicons for both models.

Sentiment analysis using Bing and NRC lexicons was done for both models shows all the topic whether at model k=3 or k=6 shows higher positive sentiments compared to negative sentiments. In the k=3 models, Topic 3 hsows the strongest sentiment values which is at 0.0803 with lowest negative sentiment at 0.0154. Topic 1 and 2 have higher positive proportions suggesting all three topics contained more favourable or encouraging content overall. For example, Topic 1 talk about double-pair badminton in Malaysia won the Open India 2025 which shows more positive compare to negative.

Then, in the k=6 models, Topic 3 again have a higher positive sentiment value at 0.0925 which much more higher than k=3 model for Topic 3. It also have the lowest negative sentiment at k=6 models at 0.0499 hints a slightly more critical or challenging tone. However there are some topic that might have equal negative and positive sentiment values which is Topic 2. This is because Topic 2 talk about the injuries and healing theme. These result shows dataset carrier a generally a positive language profile.

Next, when using NRC lexicons it captures ten sentiments for specific emotions for each topics to allow deeper emotional profiling. In the k=3 models, ‘trust’ and ‘positive’ emotions have dominate all three topics especially on Topic 3 with value 0.0891 and 0.137. This is show all the news for all topic have a positive sentiment and a good news for Malaysian.

'Anticipitation' and 'joy' also prominent while 'anger', 'disgust' and 'sadness' are low which suggesting the content are largely constructive, motivational and future-oriented.

Then, for k=6 models, shows the topic 1, 3 and 4 have a strong 'positive' and 'trust' emotions with Topic 3 again standouts with hight positive emotion like 'trust' at 0.0948. On the other hand, Topic 2 have the highest negative emotions compared to others like 'anger' at 0.0384, 'fear' at 0.0513 and 'negative' sentiments at 0.0777. This indicate a more intense or critical theme which is Topic 2 about the injuries of the athletes. Overall NRC result support the Bing findings and provide further insights into emotional makeup for each topic. Figure 7 below shows the top TF-IDF terms by topic for both models.

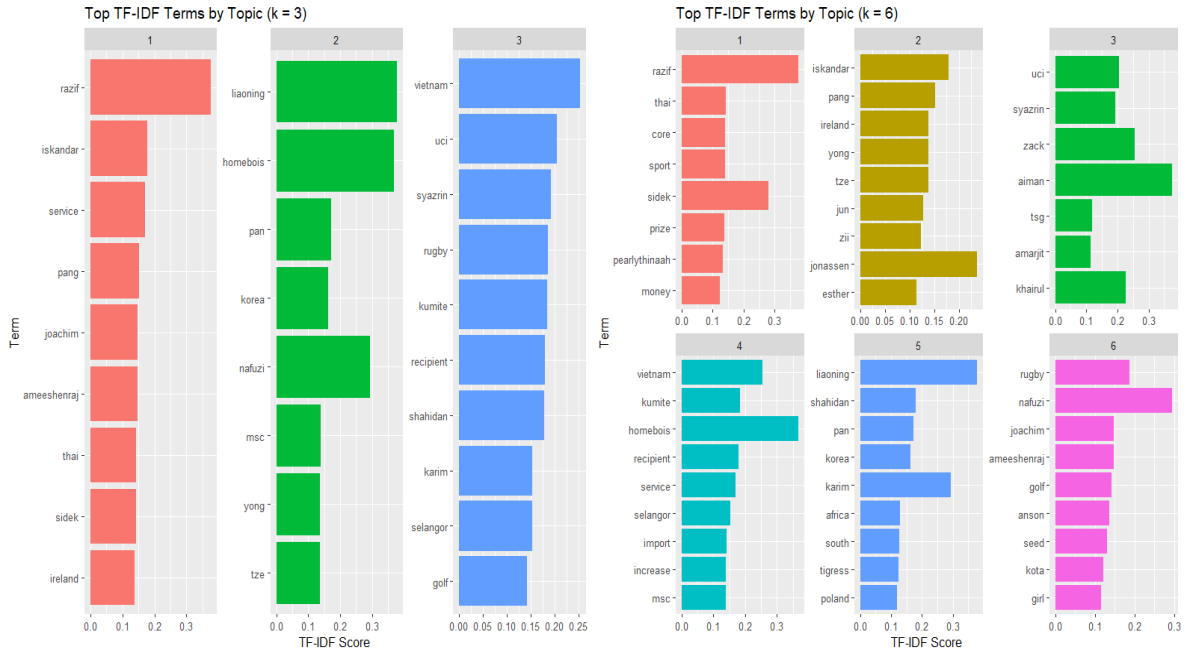

Figure 7 Top TF-IDF terms by topic for both models.

The TF-IDF reveals each topic captures a set of distinctive and high-impact terms across different document. In model k=3, Topic 1 have higher terms like name such as "Razif" and 'Iskandar'. This shows the topic might focused on the athelets or personal achievements in sports. Topic 2 is hinting at geographical references and teams such as "Homebois" and "Korea" while Topic 3 highlights sports or event such as "Golf". These TF-IDF terms not only differentiate the topics but also highlight the uniqueness of documents assigned to each cluster.

Then, for k= 6 models analysis reveals more specific subtopics. Topic 1 and Topic 2 is focused on athlete achievements and have higher terms focused on names like “Sidek” and “Iskandar”. As know Topic 1 is about badminton and the person who is name Sidek is very popular in the badminton sport in Malaysia due to his contributions. Topic 3 and 4 captures terms like “Homebois” and “Selangor” which indicates local teams or competitions. Lastly Topic 5 and Topic 6 is included international terms like “Korea” and further separating regional from global themes. The increase int topic number allows finer differentiation. Figure 8 below shows topic proportions per document for both models.

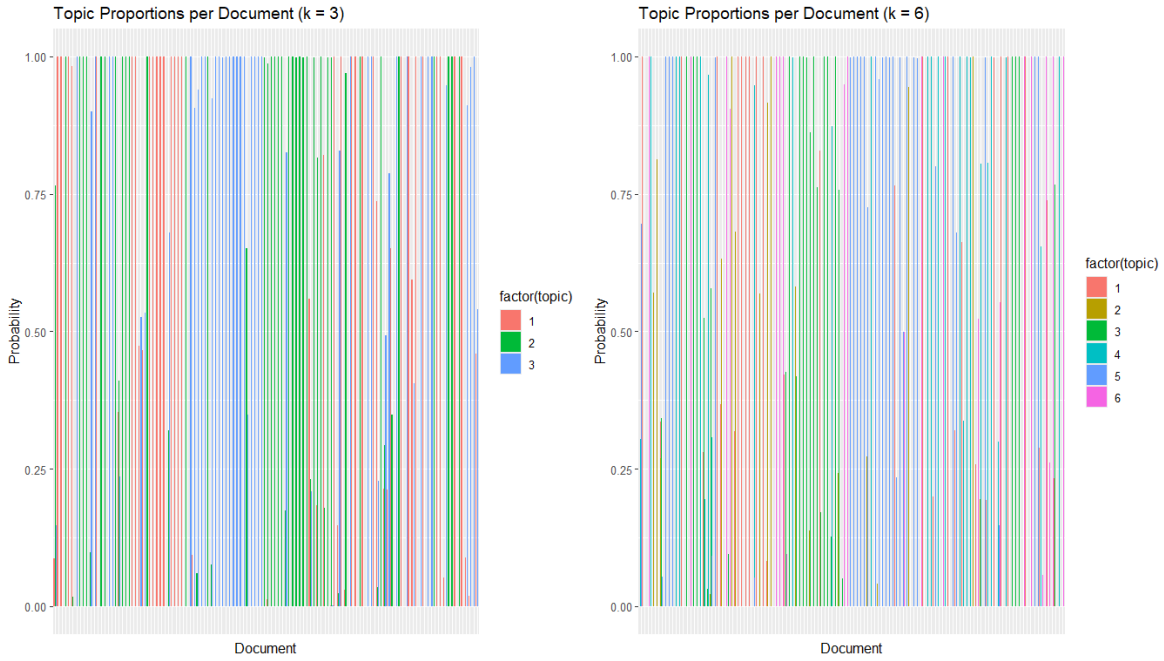

Figure 8 Topic proportions per document for both models.

The gamma is represent topic proportions for each document. For example if a document has gamma = 0.78 for Topic 2 and 0.28 for Topic 1, it means the document will fall under Topic 2. The probability is between 0 and 1 sum to one document. Figure above just to show values visually for dominant topic for each document. This also to understand document in terms of topic distributions.

Analysis and Discussion

In conclusion, the sentiment and topic modelling analysis using both k=3 and k=6 topic provide valuable insights into structure and emotion in the sport news. Both models reveal meaningful insight but k=6 model offered clearer and more focused topic which helps differentiate between sports types like badminton, football and cycling. Sentiment analysis using Bing and NRC lexicons shows all the topics carriers a positive sentiment for both models and also have higher positive emotions. Gamma values showed that most documents aligned strongly with one dominant topic, suggesting good model coherence. TF-IDF highlighted key terms that reinforced and clarified the topics identified by LDA. All the analysis is needed to understand more about the data and get a good insight.