Projekt 2 Opinion Mining

DATA MINING

Autorzy:

Emil Filipowicz, Karolina Grulkowska, Maria Sobocińska

Analityka Gospodarcza II, semestr 1

Wprowadzenie

Badanie opinii odbiorców stanowi jedno z kluczowych źródeł informacji dla przedsiębiorców, firm, inwestorów, polityków, artystów oraz dla wszystkich tych, których branża oraz praca zależy od opinii publicznej. Dlatego też w wielu dziedzinach ważne jest, aby przeprowadzać odpowiednie analizy opinii dotyczących danych produktów, utworów czy firm. Informacja stanowi w tym przypadku wiodący punkt wyjścia. Interesariusze oraz podmioty otrzymując odpowiednie komunikaty z rynku, są w stanie podjąć odpowiednie kroki w zależności od warunków oraz wprowadzić strategię umożliwiającą poprawienie sytuacji lub wykorzystanie nadarzających się szans, rozwijając lub poprawiając dane produkty czy dostosowując się do określonych warunków rynkowych.

W poniższej pracy przedstawiona została analiza opinii klientów różnych modeli telefonów, pochodzących od 5-ciu globalnych oraz popularnych w 2020 roku producentów smartphone'ów. Do realizacji projektu zostały użyte metody oraz narzędzia Text Mining 'u, które są jedną z gałęzi większej dziedziny o nazwie Data Mining. Marki oraz modele urządzeń na podstawie, których zostało przeprowadzone badanie to:

- Apple – iPhone SE 2020
- XIAOMI - Redmi Note 9 Pro
- HUAWEI - P40 Lite
- SAMSUNG - SM-A715 Galaxy A71
- MOTOROLA - Moto G8 Power

W początkowej fazie projektu zaprezentowane zostało szersze *Wprowadzenie* mające pozwolić na odpowiednie zapoznanie się z problematyką badania oraz z sprecyzowanym celem przeprowadzonej analizy. W kolejnym etapie przedstawiono uwarunkowania teoretyczne, metody oraz narzędzia użyte do wykonania badania. Zapoznanie się z nimi jest koniecznym punktem, aby zrozumieć sens używanych w projekcie terminów oraz zrozumieć poszczególne etapy rozpatrywanych zagadnień. Informacje te kolejno znalazły się w rozdziałach *Zarys Teoretyczny,* w którym przedstawiono fundamentalną wiedzę z zakresu Text Mining 'u oraz w *Metodologii Badań*, w której zawarte zostały specjalistyczne wyjaśnienie wspomnianych już wcześniej terminów. Po objaśnieniu teoretycznej strony projektu przedstawiony został problem badawczy oraz sekwencja metod wykorzystanych w analizie. Kolejny faza to omówienie głównego etapu eksploracji dokumentów tekstowych

zawierających opinie klientów na temat wybranych modeli smartphone'ów, użycie wymienionych wcześniej metod oraz interpretacja wyników z nich uzyskanych, pozwoliła ona również na przejście do ostatniego etapu badań. W *Podsumowaniu* zawarte zostało porównanie ostatecznie otrzymanych wyników oraz przedstawienie wniosków z przeprowadzonej analizy.

1. Zarys Teoretyczny

Data mining to proces analityczny przeznaczony do badania dużych zasobów danych w poszukiwaniu wzorców oraz współzależności pomiędzy zmiennymi. Jedną z dziedzin wywodzących się od "*zgłębiania danych*" jest text mining - eksploracja tekstu, bazująca na przetwarzaniu języka poprzez technologię sztucznej inteligencji. Oba zagadnienia są ze sobą ściśle powiązane, jako iż text mining przy użyciu odpowiednich narzędzi zajmuje się analizą tekstu, eksploracją, sporządzaniem streszczeń, podziałem na grupy dokumentów oraz wyszukiwaniem słów o podobnym znaczeniu, co następnie zostaje wykorzystane do wykrywania wzorców w danych liczbowych podczas data mining.

Przed rozpoczęciem analizy przeprowadzony zostaje preprocessing - technika używana do przekształcania surowych danych w użyteczny i wydajny format - w celu zaprezentowania jej działania na korpusie utworzonym w ramach niniejszego projektu, przytoczony został następujący komentarz:

```
> writeLines(as.character(docs [[12]]))
The phone works great, I've had it for a month. This is my first Apple product and certainly not the l
ast. The quality of photos is great, the only downside is the battery, which could hold a bit longer.
 I recommend !!!
```

Przed zrealizowaniem ostatecznej analizy, z korpusu nastąpiło usunięcie znaków specjalnych, liczb, stopwords i spacji, transformacja dużych liter na małe oraz steaming. Ograniczeniom zostały również poddane wyrazy, a przy analizie opinii uwzględniono:

- Słowa składające się od 3 do 15 liter
- Słowa, które muszą pojawiać się w co najmniej dwóch opiniach,

w wyniku czego, ten sam komentarz po transformacji prezentuje się następująco:

```
> writeLines(as.character(docs[[12]]))
work great ive month first appl product certain last qualiti photo great downsid batteri hold bit long
er recommend
```

2. Metodologia Badań

- **DATA MINING** – eksploracja danych, część informatyki, zajmująca się odkrywaniem wzorców w dużych zbiorach danych.
- **TEXT MINING** – dziedzina zajmująca się przetwarzaniem zbiorów dokumentów w celu znalezienia informacji
- **KORPUS** – określa dużą kolekcję dokumentów, opisanych i sprawdzonych np. w szczególnym przypadku do opisywanej postaci reprezentacji wektorowej
- **PRZETWARZANIE WSTĘPNE** – przekształcenie nieprzetworzonych danych w zrozumiały format.
- **PREPROCESSING** – Typowe zadania czyszczenia danych związane z eksploracją tekstu, np. usuwanie znaków interpunkcyjnych; przekształcenie „surowych" danych na zrozumiały format.
- **DOCUMENT TERM MATRIX** – rodzaj macierzy opisującej częstotliwość występowania słów (terminów) w zbiorze dokumentów należących do korpusu. W DTM dokumenty są reprezentowane przez wiersze i terminy (lub słowa) według kolumn.
- **TERM FREQUENCY** – częstotliwość wystąpienia sumy wyrazów w określonym dokumencie; częstotliwość tokena w dokumencie.
- **TOKEN** – to symbole słownictwa danego języka.
- **SPARSITY** – procent komórek (lub wierszy) w bazie danych, które są puste; służy identyfikacji średniej liczby komórek, które są rzadkie lub niewykorzystane.
- **STOP WORDS** – sposób eliminowania słów, które są nieistotne z punktu widzenia analizy korpusu; nie są nośnikiem istotnej informacji; najczęściej jest to grupa wyrazów.
- **REMOVE SPARSES** – usuwanie pustych komórek bądź wierszy w celu pozbycia się zakłóceń.
- **STRIP WHITE SPACE** – funkcja usuwająca powstałe odstępy pomiędzy wyrazami, tzw. „białe" spacje.
- **PRAWO ZIPFA** – prawo opisujące zasadę częstotliwości użycia w dowolnym języku poszczególnych wyrazów; zazwyczaj jest to 20% słów, które mają największą częstotliwość występowania, informuje o rozkładzie teoretycznym wyrazów w tekście.
- **WORD CLOUD** – prezentacja graficzna tekstu, wyrazów charakteryzujących się określoną częstotliwością w postaci chmury.
- **STEMMING** – przekształcanie wyrazów do formy podstawowej (rdzenia)
- **LEMATYZACJA** – (metoda słownikowa) – korzysta ze słownika morfologicznego i analizuje kontekst słów, zwłaszcza form słów stojących obok siebie
- **NORMALIZACJA** – proces przetwarzania tekstów, zapewniający ich porównywalność, celem ułatwienia dalszej interpretacji; proces ujednolicenia tekstów.
- **TOPIC MODELLING** – rodzaj modelowania statystycznego służący do odkrywania tematów (topiców) występujących w zbiorze dokumentów.

- **MACIERZ CZĘSTOTLIWOŚCI** – jest to rodzaj macierzy, w której przedstawiono częstotliwość występowania poszczególnych wyrazów w danym tekście.
- **SKALOWANIE WIELOWYMIAROWE** – zestaw wielowymiarowych metod analizy danych,do których się stosuje analizować podobieństwa lub rozbieżności w danych.
- **ANALIZA SENTYMENTU** - proces, który analizuje fragment tekstu, aby ocenić jego nacechowanie emocjonalne i tym samym bada wydźwięk wypowiedzi.

3. Problem oraz cel badawczy

Problem badawczy oraz określenie pytania badawczego jest kluczowym elementem każdej analizy, pomaga on wyznaczyć kierunek badania oraz dobrać odpowiednie metody pozwalające uzyskać optymalne wyniki. Niniejszy projekt opiera się na badaniu opinii klientów poszczególnych producentów wybranych smartphone'ów, porównaniu komentarzy oraz wyznaczaniu najlepszego telefonu względem zebranych opinii. Przed rozpoczęciem szczegółowej analizy zostało określone 3 pytania badawcze.

1. W jaki sposób nabywcy wybranych smartphone'ów wypowiadają się na temat swojego zakupu?
2. Które elementy smartphonów były najczęściej komentowane przez nabywców?
3. Który z wybranych modeli smartphonów uzyskał najlepszy wynik względem komentarzy?

Projekt ten oraz przeprowadzane w nim działania analityczne będą miały na celu odpowiedzieć na wyznaczone pytania i określić ogólne trendy zakupowe wśród klientów wybranych smartphonów.

4. Plan Badania

Do lepszego zrozumienia procesów zachodzących w pracy, umieszczona została sekwencja użytych metod podczas przeprowadzania analizy.

1. Pozyskanie danych w postaci opinii
2. Preprocessing korpusu
3. Analiza korpusu.
4. Topic modelling.
5. Analiza statystyczna struktury Topiców, wyznaczenie udziału procentowego przedstawiającego, ile opinii wchodzi w skład danego Topica
6. Analiza sentymentu poszczególnych producentów telefonów
7. Analiza względem wydźwięku opinii dla każdej marki. Sprawdzenie czy są pozytywne, negatywne lub neutralne.
8. Grupowanie marek na podstawie wspólnych topic'ów
9. Ocena i analiza wyników

5 Analiza wyników

5.1 Wprowadzenie do analizy

Rozdział ten poświęcony został analizie, porównaniu oraz podsumowaniu wyników, powstałych podczas badania opinii na temat 5-ciu topowych modeli smartphonów. Obserwacje zostały przeprowadzone oraz uzyskane za pomocą środowiska R.
Podczas procesu generowania niezbędnych informacji do uzyskania wyników zostały użyte następujące biblioteki:

- library(ggplot2)
- library(wordcloud)
- library(topicmodels)
- library(stringr)
- library(syuzhet)
- library(tm)
- library(plyr)

Po wprowadzeniu pliku csv. zawierającego opinie na temat wybranych modeli do środowiska R został utworzony DTM dla całej bazy dokumentów.

```
> dtm <- DocumentTermMatrix(docs)
> dtm
<<DocumentTermMatrix (documents: 250, terms: 1546)>>
Non-/sparse entries: 4912/381588
Sparsity           : 99%
Maximal term length: 16
Weighting          : term frequency (tf)
```

W wyniku podjętego zabiegu powstała macierz o wymiarach 250 na 1546, gdzie pierwsza wartość oznacza liczbę dokumentów w korpusie, a druga zaś ilość słów w nim występujących. Gdzie rozproszenie macierzy (sparsity) wynosi 99%, a najdłuższe słowo ma długość 16-tu liter. W kolejnym etapie eksploracji korpusu z opiniami, zostały wyselekcjonowane poszczególne modele smartphonów oraz zostały poddane indywidualnej obróbce, analizie oraz ocenie.

5.2 APPLE - iPhone SE 2020

W pierwszej kolejności badaniu został poddany produkt firmy Apple - IPhone SE 2020, gdzie zestawionych zostało 50 opinii zakupowych odnośnie wybranego telefonu. W tym celu nastąpiła transformacja pliku na ramkę danych poprzez selekcje obserwacji dla wcześniej wspomnianego producenta. Dzięki temu powstała nowa macierz DTM zawierająca w sobie informacje odnośnie nowo utworzonego obiektu.

```
> dtm <- DocumentTermMatrix(docs)
> dtm
<<DocumentTermMatrix (documents: 50, terms: 525)>>
Non-/sparse entries: 999/25251
Sparsity           : 96%
Maximal term length: 15
Weighting          : term frequency (tf)
```

DTM powstały dla APPLE – iPhone dysponuje wymiarami w liczbie 50 opinii na 525 słów. Gdzie rozproszenie macierzy (sparsity) wynosi 96%, a najdłuższe słowo ma długość 15-tu liter.

5.2.1 Statystyki opisowe dla opinii – Apple

- Opinia o maksymalnej długości

```
> max_length<-max(doc_length)
> max_length
[1] 102
```

- Opinia o minimalnej długości

```
> min_length<-min(doc_length)
> min_length
[1] 2
```

- Średnia długość opinii w korpusie

```
> aver_length<-mean(rowSums(as.matrix(dtm)))
> aver_length
[1] 23.28
```

5.2.2 Preprocessing

Przykładowa opinia przed preprocessingiem:

```
> writeLines(as.character(docs [[12]]))
The phone works great, I've had it for a month. This is my first Apple product and certainly not the l
ast. The quality of photos is great, the only downside is the battery, which could hold a bit longer.
 I recommend !!!
```

Opinia po przeprowadzeniu preprocessingu:

```
> writeLines(as.character(docs[[12]]))
work great ive month first appl product certain last qualiti photo great downsid batteri hold bit long
er recommend
```

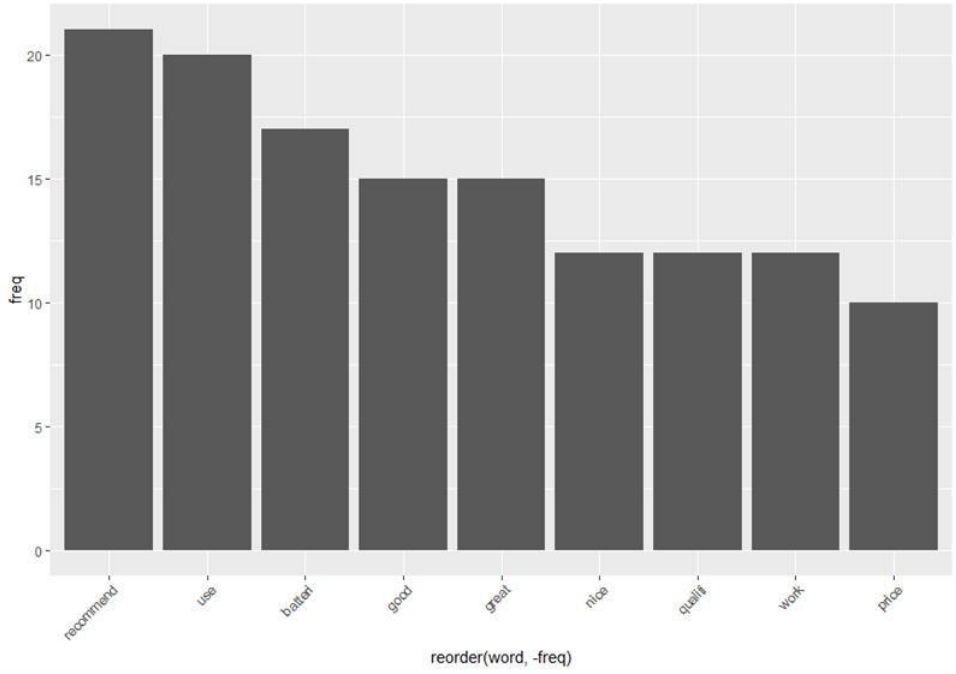
5.2.3 DTM dla opinii po przeprowadzeniu procesu preprocessingu

```
<<DocumentTermMatrix (documents: 50, terms: 112)>>
Non-/sparse entries: 430/5170
Sparsity           : 92%
Maximal term length: 11
Weighting          : term frequency (tf)
```

DTM powstały dla APPLE – iPhone preprocessingu dysponuje wymiarami w liczbie 50 opinii na 112 słów, gdzie rozproszenie macierzy (sparsity) wynosi 92%, a najdłuższe słowo ma długość 11-tu liter.

5.2.4 Zipf

Graficzne przedstawienie najczęściej występujących słów za pomocą histogramy opartego na prawie Zipfa.



Najczęstszym słowem pojawiającym się w korpusie było “recommend”, następnie ”use” (od “useful”) oraz “batteri”. Częstotliwość występowania tych terminów pokazało pozytywny odbiór smartphona wśród nabywców oraz świadczy o tym, że klienci dzięki wydajności jaką oferuję telefon polecają go do zakupu. Kolejne słowa potwierdziły tę tezę.

5.2.5 Wordcloud

Wordcloud jest kolejną metodą graficznego przedstawienia częstotliwości występowania słów.

Powyższe Wordcloudy potwierdzają wynik osiągnięty w pierwszej metodzie graficznej. Biorąc pod uwagę przedstawione chmury słów, klienci w swoich opiniach wskazywali na wydajność oraz możliwości telefonu, dzięki czemu chętnie polecali go i pisali o nim w superlatywach.

5.2.6 Topic Modelling

Przy użyciu metody Topic Modelling'u stworzone zostały główne tematy zawierające najistotniejsze słowa występujące w badanym korpusie. W tabeli natomiast przedstawiono nazwy oraz opisy poszczególnych topic'ów, liczbę opinii oraz ich numery, które końcowo złożyły się na zawartość poszczególnych tematów. Następnie przedstawiono scoring oraz analizę opinii w kolejnych topicach.

```
> ldaOut.terms
     Topic 1 Topic 2     Topic 3  Topic 4
[1,] "use"   "recommend" "work"   "batteri"
[2,] "nice"  "qualiti"   "day"    "great"
[3,] "appl"  "camera"    "good"   "good"
[4,] "price" "smooth"    "realli" "product"
[5,] "last"  "alway"     "better" "high"
```

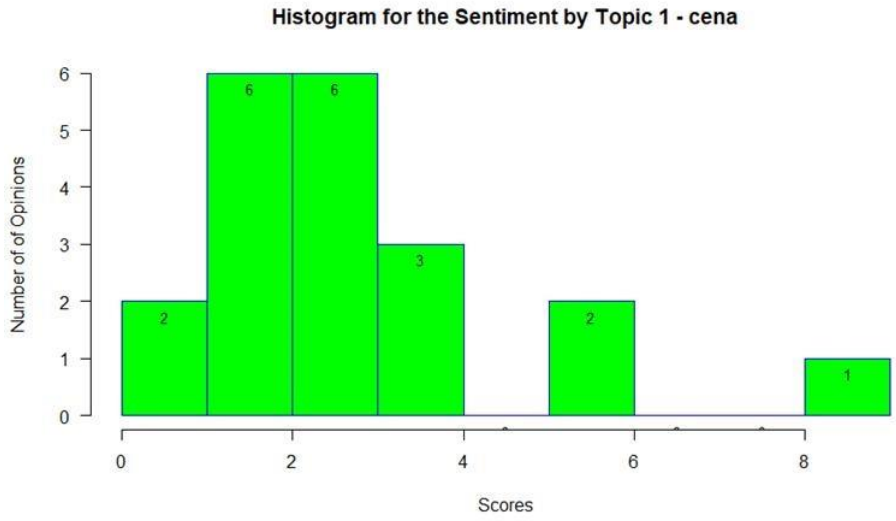
	TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4
NAZWA	CENA	KAMERA	UŻYTKOWANIE	BATERIA
OPIS	Cena w stosunku do użyteczności urządzenia firmy Apple	Polecenie smartphona ze względu na jakość kamery	Codzienne użytkowanie urządzenia	Pozytywne reakcje klientów w stosunku do wysokich możliwości baterii
LICZBA OPINII	20	15	8	7
NUMER KOMENTARZA	5, 7, 9, 13, 14, 20, 23, 25, 27, 28, 29, 30, 31, 37, 39, 41, 44, 48, 49, 50	1, 3, 6, 10, 11, 15, 16, 21, 24, 32, 33, 36, 38, 40, 42,	18, 19, 22, 26, 35, 43, 45, 47,	2, 4, 8, 12, 17, 34, 46

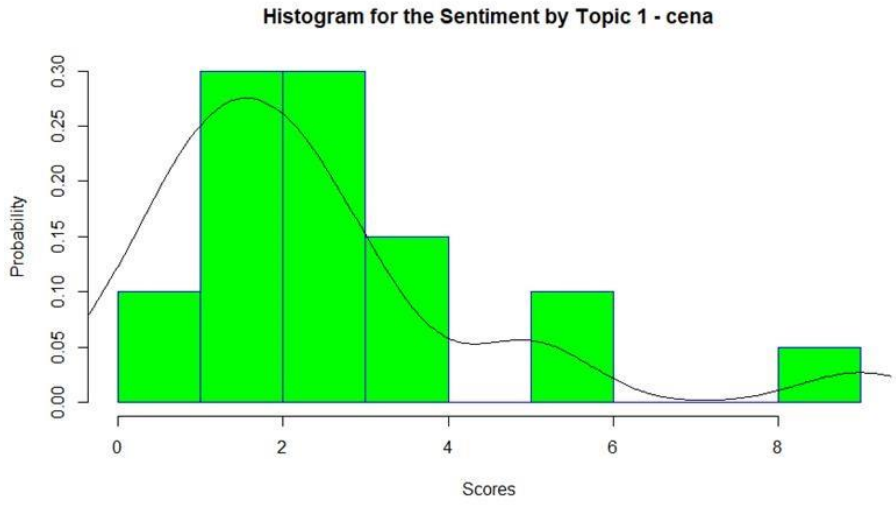
Jak możemy odczytać z tabelki, aż 40% opinii dotyczy ceny w stosunku do użyteczności produktu. Najmniej osób zaś wypowiadało się na temat możliwości baterii, recenzje te stanowią tylko 14% wszystkich opinii.

TOPIC 1 – Cena

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     1.0     2.0     2.3     3.0     9.0
> minn<-min(m1$Score)
> minn
[1] 0
> maxx<-max(m1$Score)
> maxx
[1] 9
```






```
> pos1$Score
 [1] 2 1 3 2 1 3 1 2 2 9 5 2 2 3 1 1 5 1
> length(pos1$Score)
[1] 18
```

```
> neu1$Score
[1] 0 0
> length(neu1$Score)
[1] 2
```

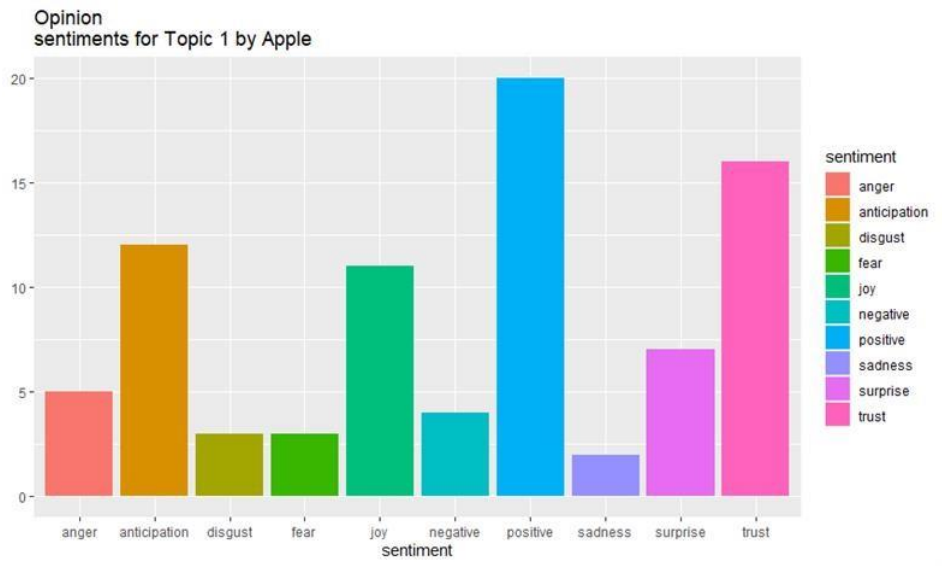
```
> neg1$Score
integer(0)
> length(neg1$Score)
[1] 0
```

Zebrane wyniki ze scoringu dla topicu 1 posiadają pozytywny wydźwięk. Na temat ceny pozytywnie wypowiedziało się 18 nabywców urządzenia, 2 osoby pozostawiły neutralne opinie. W tym przypadku nie pojawiły się opinie negatywne.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych (po lewej stronie) i neutralnych (po prawej). Można z nich wnioskować, że klienci najbardziej cenili sobie wydajność baterii, pracy smartphonu, ekranu, oraz stosunek ceny do jakości, dzięki czemu chętnie rekomendowali go i pisali o nim w superlatywach. Opinie neutralne również wskazują na pozytywne odczucia, skupiając się ostatecznie na ogólnym zadowoleniu z zakupu.

- **Opinion sentiments**



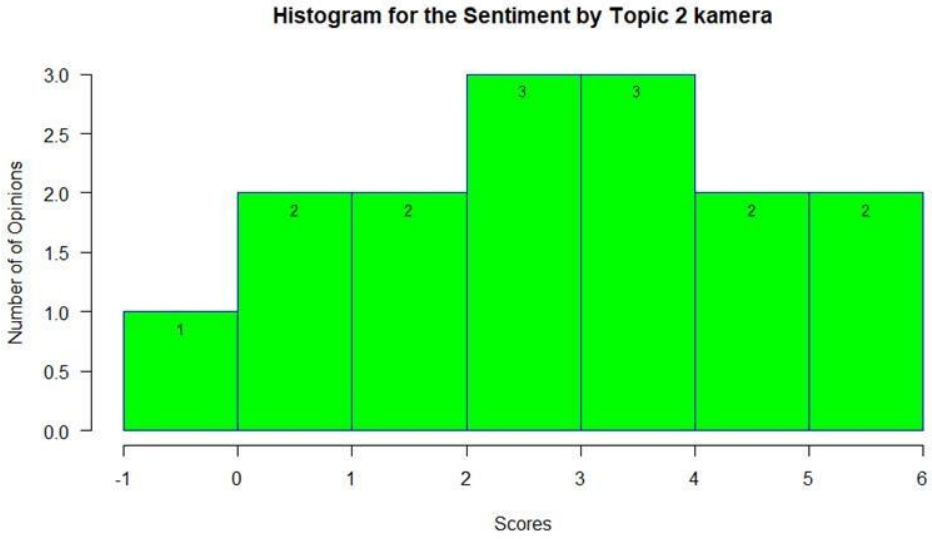
Powyższy wykres przedstawia badanie opinii klientów pod względem ich odczuć oraz emocji przypisanych do słów. Program przypisuje do określonych słów w korpusie odczucia pozytywne oraz negatywne na podstawie specjalnie do tego przeznaczonych leksykonów. Uczucia opisane na histogramie to od lewej: gniew, oczekiwania, rozczarowanie, strach, radość, negatyw, pozytyw, smutek, zaskoczenie oraz zaufanie. Główne wskaźniki osiągające

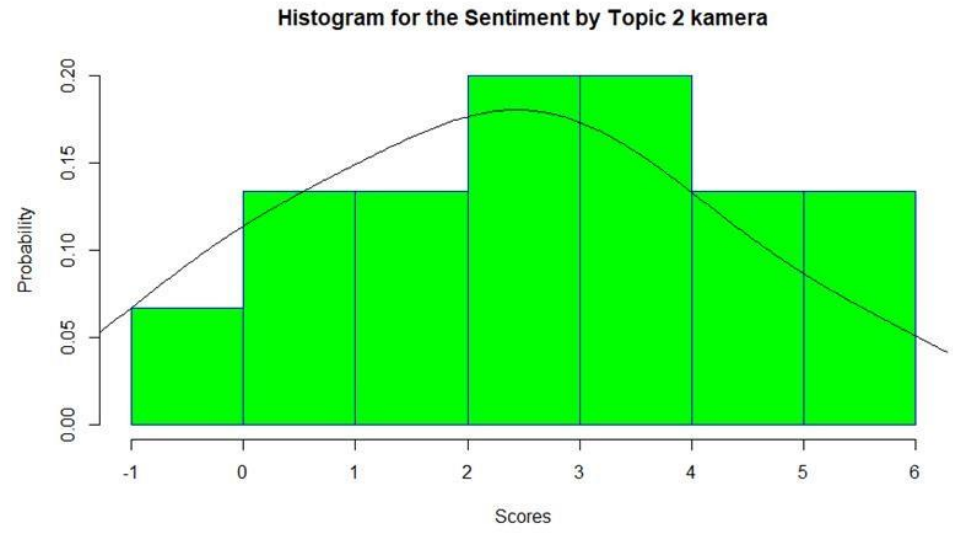
najwyższe wartości w tym przypadku to kolejno pozytyw, zaufanie i oczekiwania. Najmniejsze wartości zostały przypisane takim uczuciom jak smutek, strach oraz rozczarowanie.

TOPIC 2 - Kamera

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.000   1.000   2.000   2.333   3.500   6.000
> minn<-min(m1$Score)
> minn
[1] -1
> maxx<-max(m1$Score)
> maxx
[1] 6
```





```
> pos1$Score
 [1] 6 4 2 2 4 2 1 3 1 3 5 3
> length(pos1$Score)
[1] 12
```

```
> neu1$Score
[1] 0 0
> length(neu1$Score)
[1] 2
```

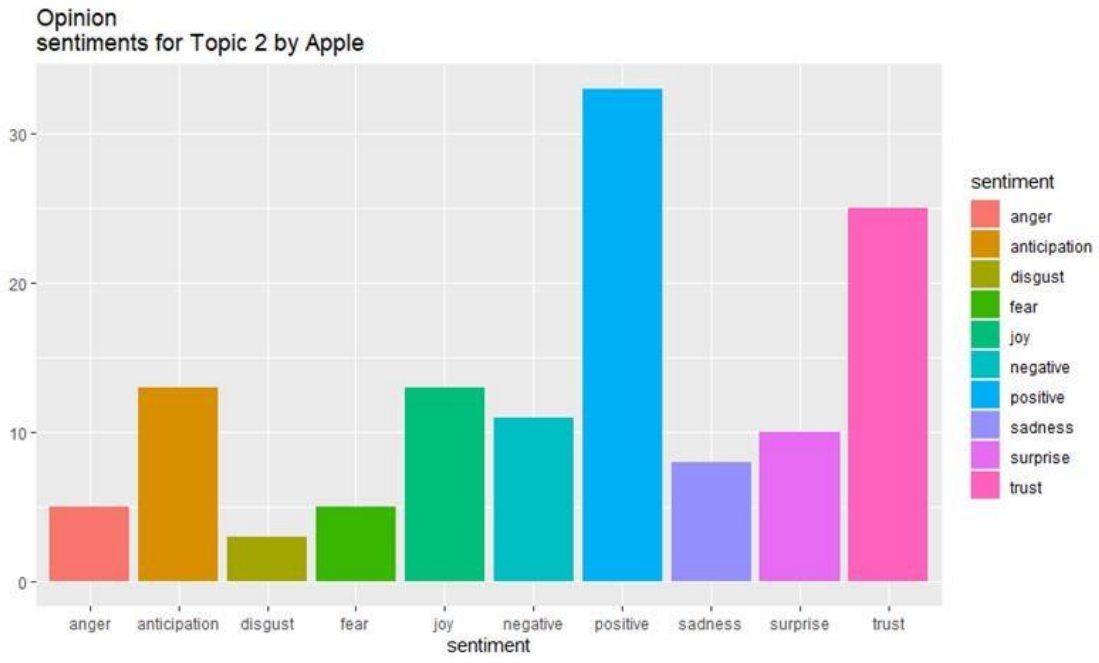
```
> neg1$Score
[1] -1
> length(neg1$Score)
[1] 1
```

Zebrane wyniki ze scoringu dla topic'u 2 posiadają wydźwięk w większości pozytywny. Na temat Kamery pozytywnie wypowiedziało się 12 nabywców urządzenia, 2 osoby pozostawiły neutralne opinie oraz pojawiła się 1 opinia negatywna.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych (po lewej stronie), neutralnych (pośrodku) i negatywnych (po prawej). Można z nich wnioskować, że klienci najbardziej cenili sobie wysoką jakość kamery i wykonywanych zdjęć, baterii oraz samego smartphonu. Opinie neutralne skupiły się na ogólnych cechach ostatecznie pozytywnych, podczas gdy jedyny komentarz negatywny zawierał w sobie informację o rozczarowaniu produktem.

- **Opinion sentiments**

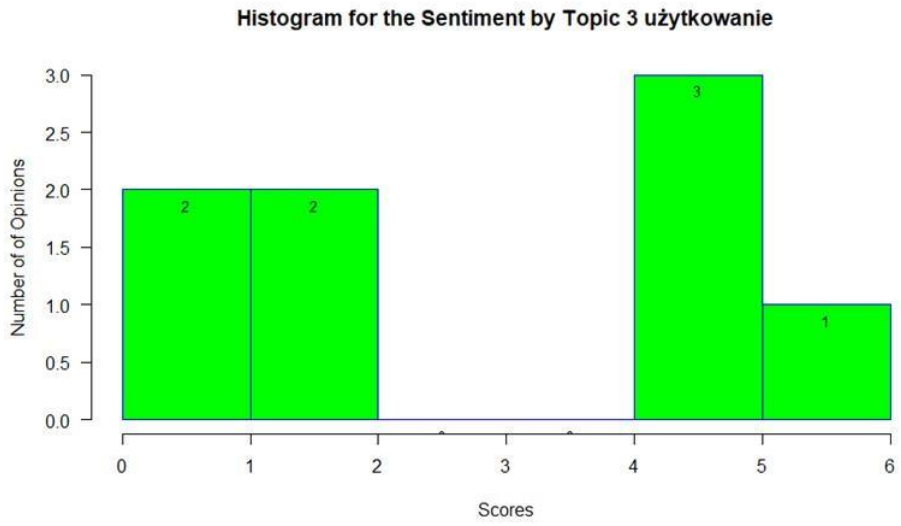


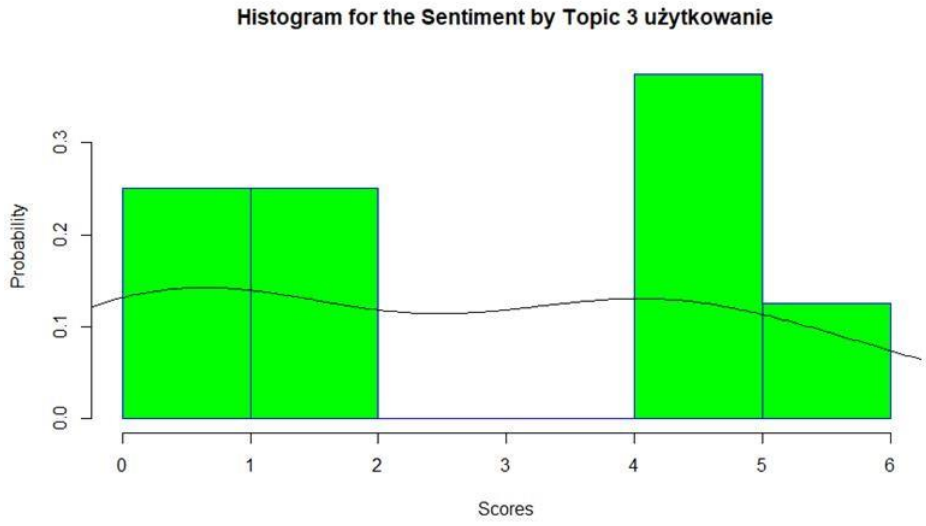
W przypadku określania sentymentu dla topic'u 2 największe wartości osiągnęły odczucia pozytywne, zauważalne było także zadowolenie klientów oraz zaufanie jakim darzą producenta nabytego urządzenia. Najmniejsze wartości osiągnęły negatywne odczucia, w małym ułamku w opinii można było dostrzec także smutek czy gniew.

TOPIC 3 - Użytkowanie

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.75    2.50    2.50    4.00    6.00
> minn<-min(m1$Score)
> minn
[1] 0
> maxx<-max(m1$Score)
> maxx
[1] 6
```





```
> pos1$Score
[1] 6 4 4 4 1 1
> length(pos1$Score)
[1] 6
```

```
> neu1$Score
[1] 0 0
> length(neu1$Score)
[1] 2
```

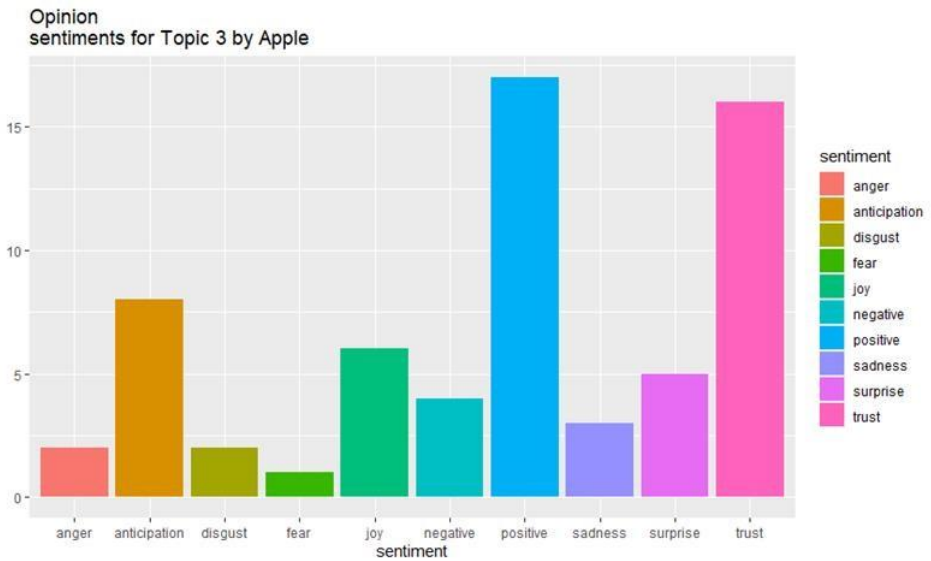
```
> neg1$Score
integer(0)
> length(neg1$Score)
[1] 0
```

Kolejny raz wyniki osiągnięte ze scoringu tym razem dla topic'u 3 posiadają wydźwięk głównie pozytywny. Na temat użytkowania smartphona pozytywnie wypowiedziało się 6 nabywców, 2 osoby pozostawiły neutralne opinie na jego temat oraz nie pojawiły się negatywne opinie.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych (po lewej stronie) i neutralnych (po prawej). Można z nich wnioskować, że klienci najbardziej cenili sobie design oraz wysoki komfort użytkowania i korzystania ze smartphonu, uważając go za godnego polecenia. Opinie neutralne skupiły się na ogólnych cechach, również w kategorii komfortu użytkowania produktu.

- **Opinion sentiments**



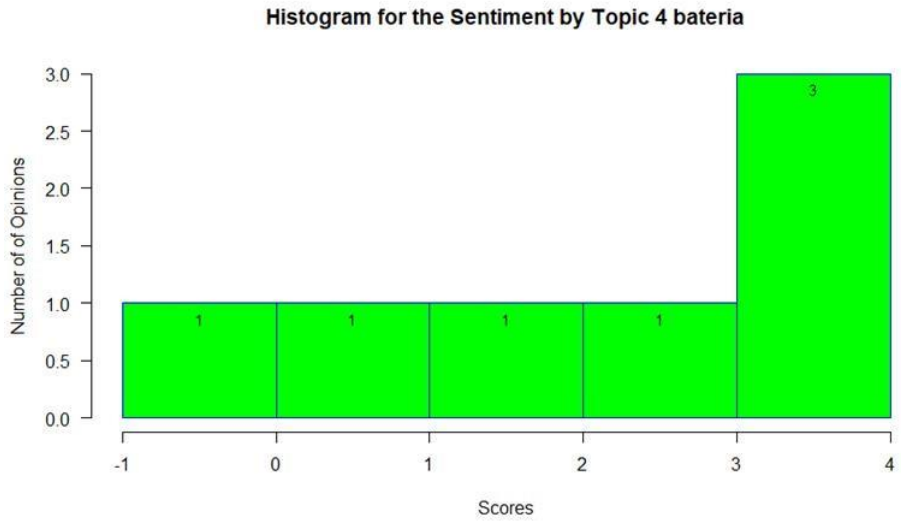
Sentyment dla topic'u 3 wskazał, że najwięcej opinii w tym temacie oscyluje na poziomie emocji silnie pozytywnych. Zauważalne jest to głównie przy dwóch słupkach histogramu,

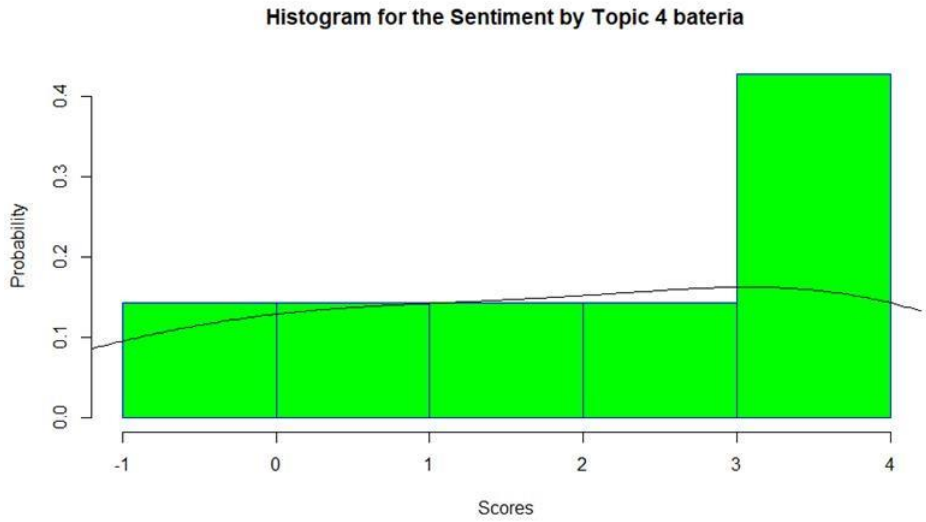
niebieskim – pozytyw oraz w kolorze magenty – zaufanie. Negatywne odczucia jak w poprzednich przypadkach charakteryzują się niskimi wartościami.

TOPIC 4 – Bateria

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.000   0.500   2.000   1.857   3.500   4.000
> minn<-min(m1$Score)
> minn
[1] -1
> maxx<-max(m1$Score)
> maxx
[1] 4
```





```
> pos1$Score
[1] 4 1 4 3 2
> length(pos1$Score)
[1] 5
```

```
> neu1$Score
[1] 0
> length(neu1$Score)
[1] 1
```

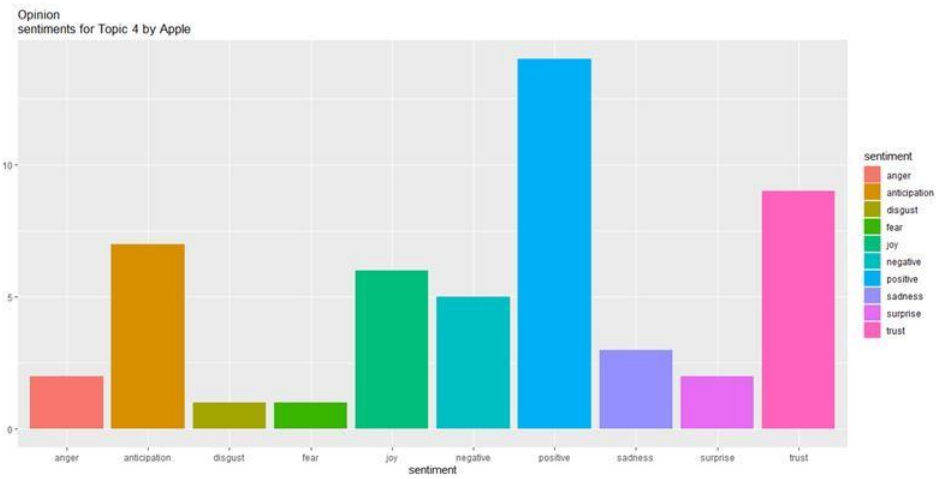
```
> neg1$Score
[1] -1
> length(neg1$Score)
[1] 1
```

Zebrane wyniki ze scoringu dla topic'u 4 posiadają wydźwięk pozytywno - neutralny. Na temat Baterii pozytywnie wypowiedziało się 5 nabywców urządzenia, 1 osoba pozostawiła neutralną opinie oraz pojawił się 1 negatywny komentarz.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych (po lewej stronie), neutralnych (pośrodku) i negatywnych (po prawej). Można z nich wnioskować, że klienci najbardziej cenili sobie wysoką wydajność baterii. Opinie neutralne skupiły się na porównaniu produktu APPLE z konkurencyjnymi odpowiednikami (będącymi smartphonami z oprogramowaniem Android), podczas gdy jedyny komentarz negatywny zawierał w sobie informację o rozczarowaniu produktem, który nie spełniał oczekiwanych wymagań.

- **Opinion sentiments**



W ostatnim przypadku dla topic'u 4 opierającego się głównie na wydajności baterii można zauważyć więcej pozytywnych reakcji. Na wysokim poziomie znajdują się także zaufanie,

oczekiwania oraz zadowolenie. Negatywy nadal nisko w porównaniu do pozytywnych odczuć mimo pojawienia się opinii negatywnej.

5.3 XIAOMI - Redmi Note 9 Pro

DTM powstały dla XIAOMI Redmi Note 9 Pro dysponuje wymiarami w liczbie 50 opinii na 442 słów, gdzie rozproszenie macierzy (sparsity) wynosi 97%, a najdłuższe słowo ma długość 14-tu liter.

```
> dtm
<<DocumentTermMatrix (documents: 50, terms: 442)>>
Non-/sparse entries: 747/21353
Sparsity           : 97%
Maximal term length: 14
Weighting          : term frequency (tf)
```

5.3.1 Statystyki opisowe dla opinii - XIAOMI

- Opinia o maksymalnej długości
```
> max_length<-max(doc_length)
> max_length
[1] 292
```
- Opinia o minimalnej długości
```
> min_length<-min(doc_length)
> min_length
[1] 1
```
- Średnia długość opinii w korpusie
```
> aver_length<-mean(rowSums(as.matrix(dtm)))
> aver_length
[1] 17.86
```

5.3.2 Preprocessing

- Przykładowa opinia przed preprocessingiem

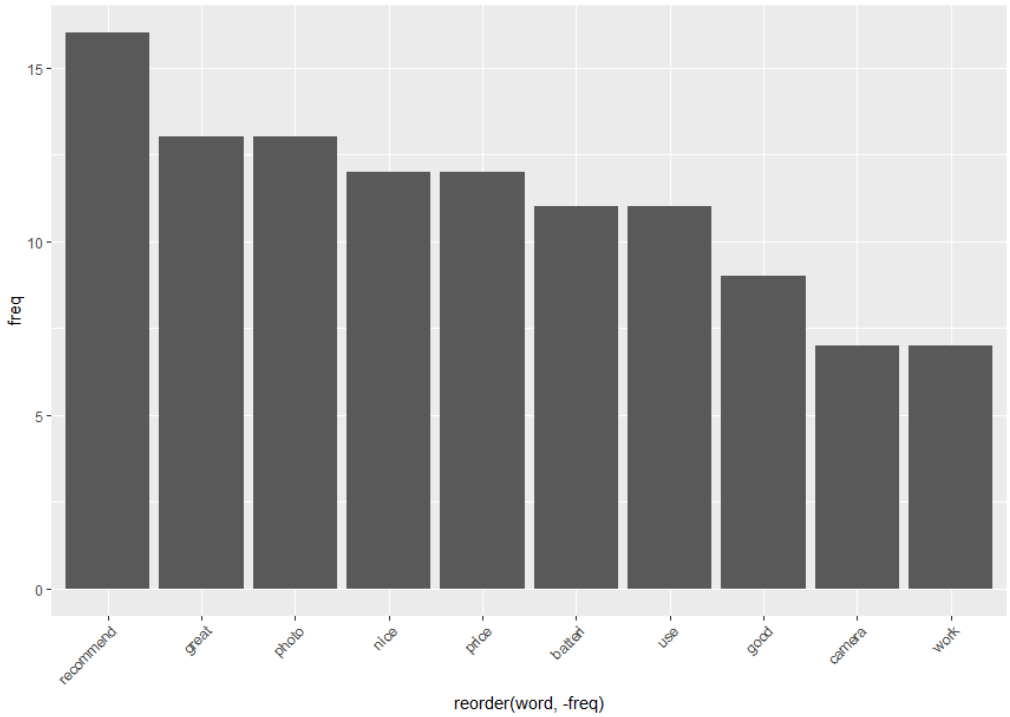
```
> writeLines(as.character(docs [[11]]))
The fingerprint reader works badly, sometimes it does not go around. NFC tested in 4 app banks hardly w
orks at al
```
- Opinia po przeprowadzeniu preprocessingu:
```
> writeLines(as.character(docs[[11]]))
fingerprint reader work bad sometim go around nfc test app bank hard work al
```

5.3.3 DTM dla opinii po przeprowadzeniu procesu preprocessingu

```
> dtm
<<DocumentTermMatrix (documents: 50, terms: 86)>>
Non-/sparse entries: 286/4014
Sparsity           : 93%
Maximal term length: 11
Weighting          : term frequency (tf)
```

DTM powstały dla urządzenia Xiaomi po preprocessingu dysponuje wymiarami w liczbie 50 opinii na 86 słów, gdzie rozproszenie macierzy (sparsity) wynosi 93%, a najdłuższe słowo ma długość 11-tu liter.

5.3.4 Zipf



Najczęstszym słowem pojawiającym się w korpusie było “recommend”, następnie ”great” oraz “photo”. Częstotliwość występowania tych terminów pokazała pozytywny odbiór smartphona wśród nabywców oraz świadczy o tym, że klienci dzięki jakości oraz funkcjom (z wyróżnieniem aparatu) jaką oferuje telefon polecają go do zakupu. Kolejne słowa potwierdziły tę tezę.

5.3.5 Wordcloud

Powyższe wordcloudy potwierdzają wynik osiągnięty w pierwszej metodzie graficznej. Biorąc pod uwagę przedstawione chmury słów, klienci w swoich opiniach wskazywali na stosunek jakości do ceny oraz możliwości telefonu (w szczególności aparatu), dzięki czemu chętnie polecali go i pisali o nim w superlatywach.

5.3.6 Topic Modelling

Podobnie jak dla poprzedniego urządzenia - za pomocą metody Topic Modelling powstały główne tematy zawierające najistotniejsze słowa występujące w korpusie opinii na temat urządzenia Xiaomi.

```
> ldaOut.terms
     Topic 1  Topic 2   Topic 3
[1,] "great"  "use"     "recommend"
[2,] "camera" "batteri" "price"
[3,] "photo"  "work"    "good"
[4,] "take"   "purchas" "nice"
[5,] "long"   "day"     "larg"
```

	TOPIC 1	TOPIC 2	TOPIC 3
NAZWA	Zdjęcia	Wydajność	Polecenie
OPIS	Temat dotyczy jakości zdjęć oraz wbudowanej kamery	Zużycie baterii oraz możliwa długość pracy telefonu	Pozytywne opinie dotyczące smartfona
LICZBA OPINII	24	11	15

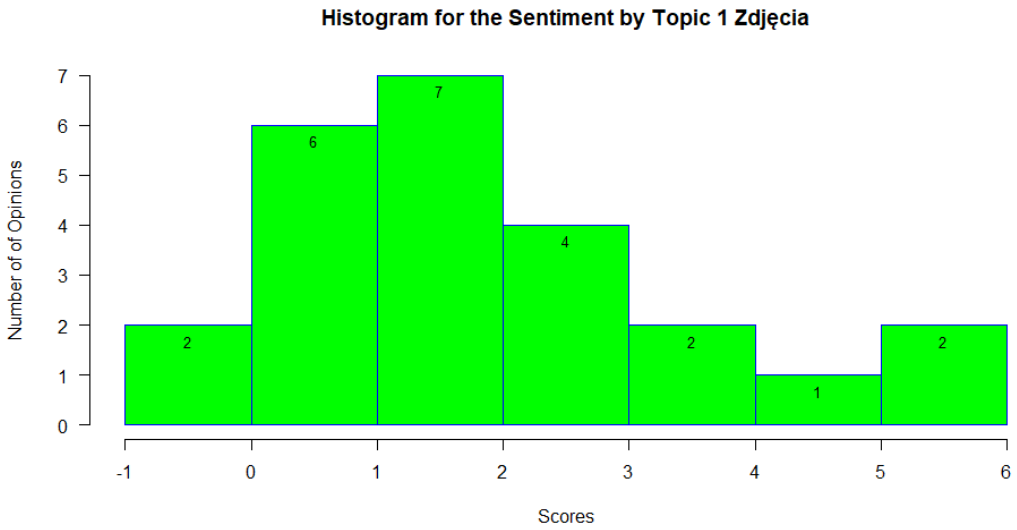
NUMER KOMENTARZA	1, 2, 3, 7, 12, 14, 15, 16, 20, 23, 25, 26, 29, 30, 36, 37, 38, 39, 41, 43, 44, 45, 46, 49	8, 11, 18, 19, 22, 24, 33, 35, 40, 42, 50	4, 5, 6, 9, 10, 13, 17, 21, 27, 28, 31, 32, 34, 47, 48

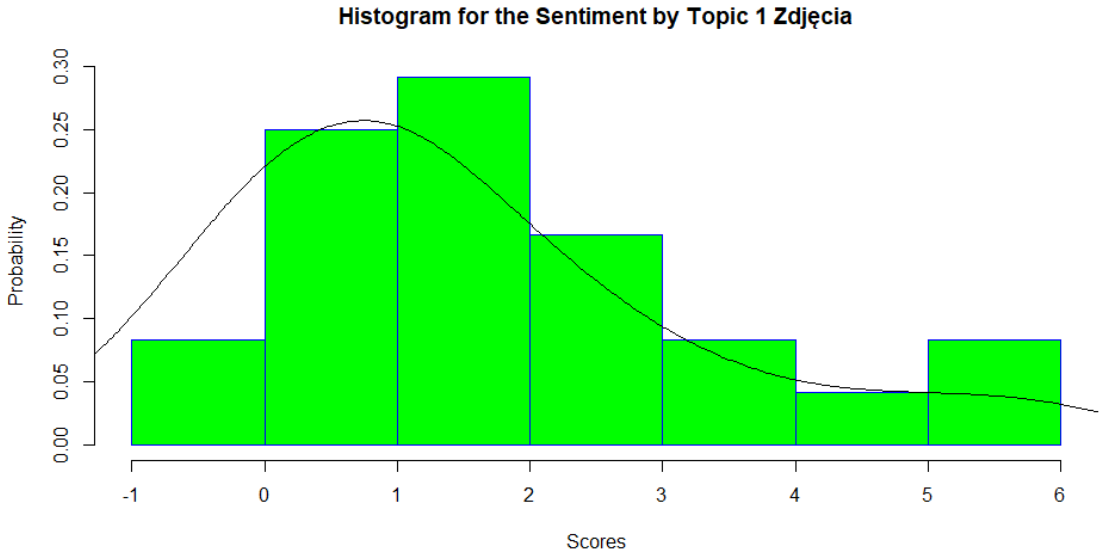
Jak wynika z powyższej tabelki znacznie wyróżniającym się tematem jest jakość zdjęć i wbudowana kamera, recenzje te stanowią aż 48% wszystkich komentarzy.

TOPIC 1 - Zdjęcia

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.000   0.000   1.000   1.417   2.000   6.000
> minn<-min(m1$Score)
> minn
[1] -1
> maxx<-max(m1$Score)
> maxx
[1] 6
```





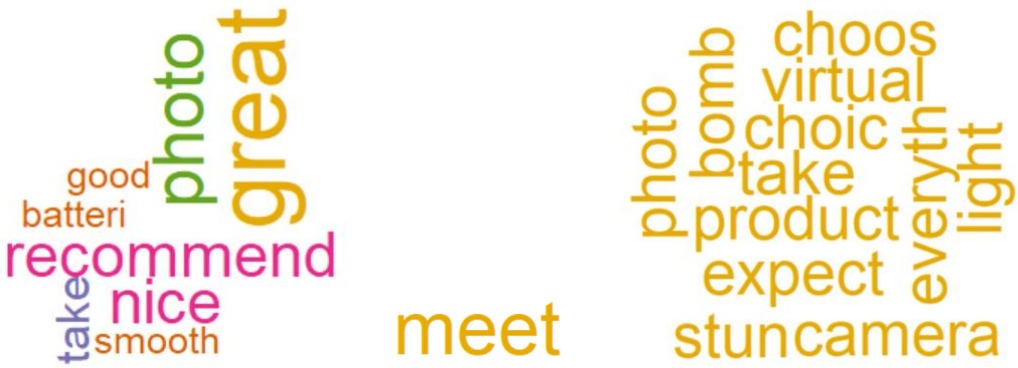

```
> pos1$Score
 [1] 3 2 2 1 1 3 1 1 1 1 2 5 2 1 4 6
> length(pos1$Score)
[1] 16
```

```
> neu1$Score
[1] 0 0 0 0 0 0
> length(neu1$Score)
[1] 6
```

```
> neg1$Score
[1] -1 -1
> length(neg1$Score)
[1] 2
```

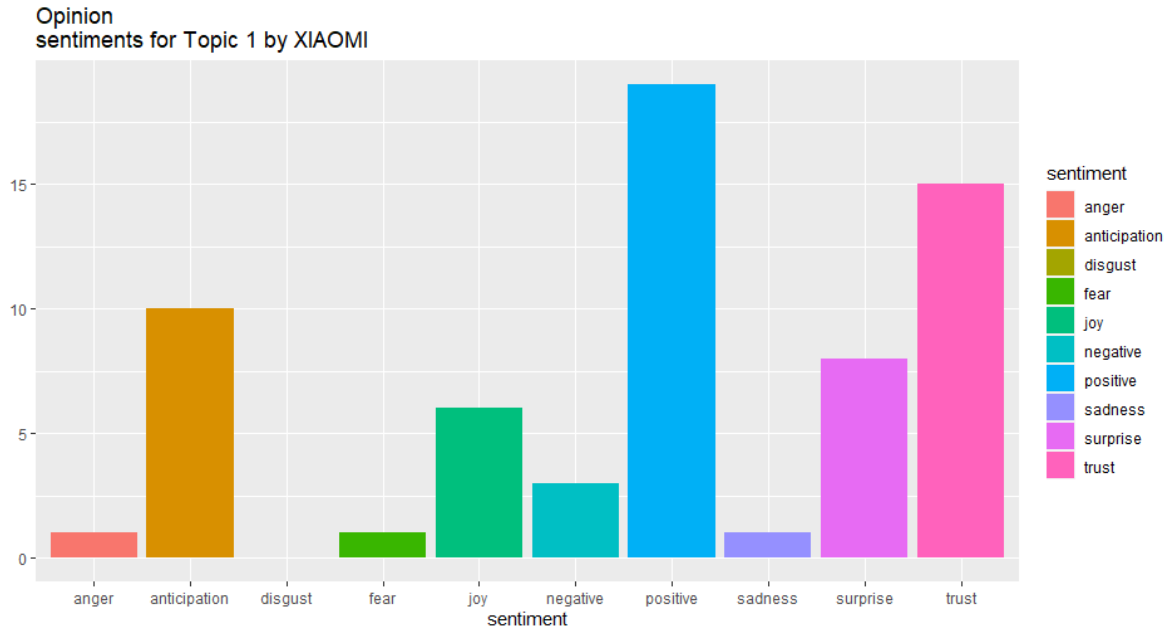
Zebrane wyniki ze scoringu dla topic'u 1 posiadają pozytywny wydźwięk. Na temat jakości zdjęć jakie robi aparat pozytywnie wypowiedziało się 16 nabywców urządzenia, 6 osoby pozostawiły neutralne opinie oraz pojawiły się dwie opinie mające kontekst negatywny.

- **Wordcloud**



Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych (po lewej stronie), neutralnych (pośrodku) i negatywnych (po prawej). Można z nich wnioskować, że klienci najbardziej cenili sobie wydajność baterii i jakość robienia zdjęć, dzięki czemu chętnie rekomendowali go i pisali o nim w superlatywach. Opinie neutralne wskazywały na to, że telefon sprostał ich oczekiwaniom, podczas gdy negatywne głosy mówiły o chaotycznym interface produktu i problemach które występują podczas robienia zdjęć.

- **Opinion sentiments**

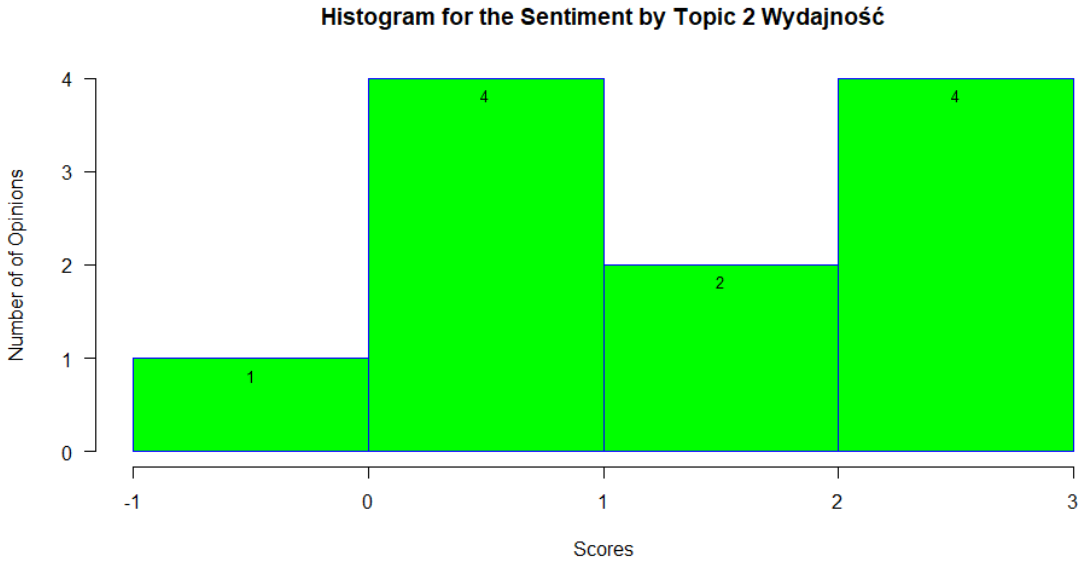


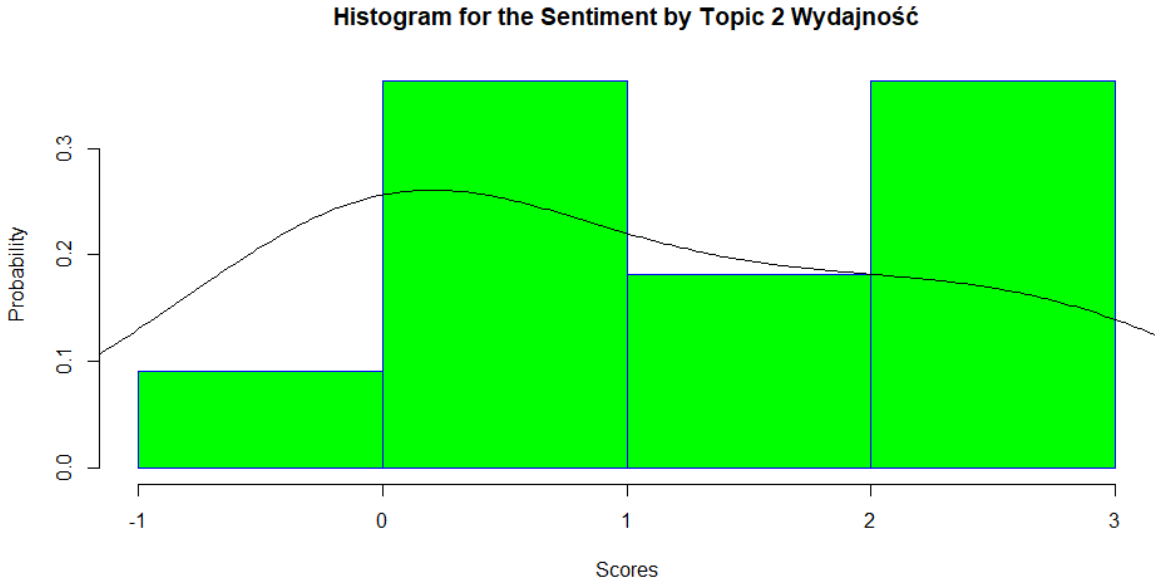
Sentyment dla topic'u 1 w Xiaomi wskazuje na pozytywne odczucia klientów odnośnie zakupu określonego modelu telefonu. Nabywcy byli zadowoleni, mile zaskoczeni oraz pozytywnie podchodzili do możliwości jakie reprezentował aparat urządzenia. Odczucia negatywne w tym przypadku są na bardzo niskim poziomie lub są całkowicie wyzerowane co wskazuję, że klienci nie rozczarowali się pod tym względem.

TOPIC 2 - Wydajność

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     -1       0       1       1       2       3
> minn<-min(m1$Score)
> minn
[1] -1
> maxx<-max(m1$Score)
> maxx
[1] 3
```






```
> pos1$Score
[1] 3 2 2 1 1 3
> length(pos1$Score)
[1] 6
```

```
> neu1$Score
[1] 0 0 0 0
> length(neu1$Score)
[1] 4
```

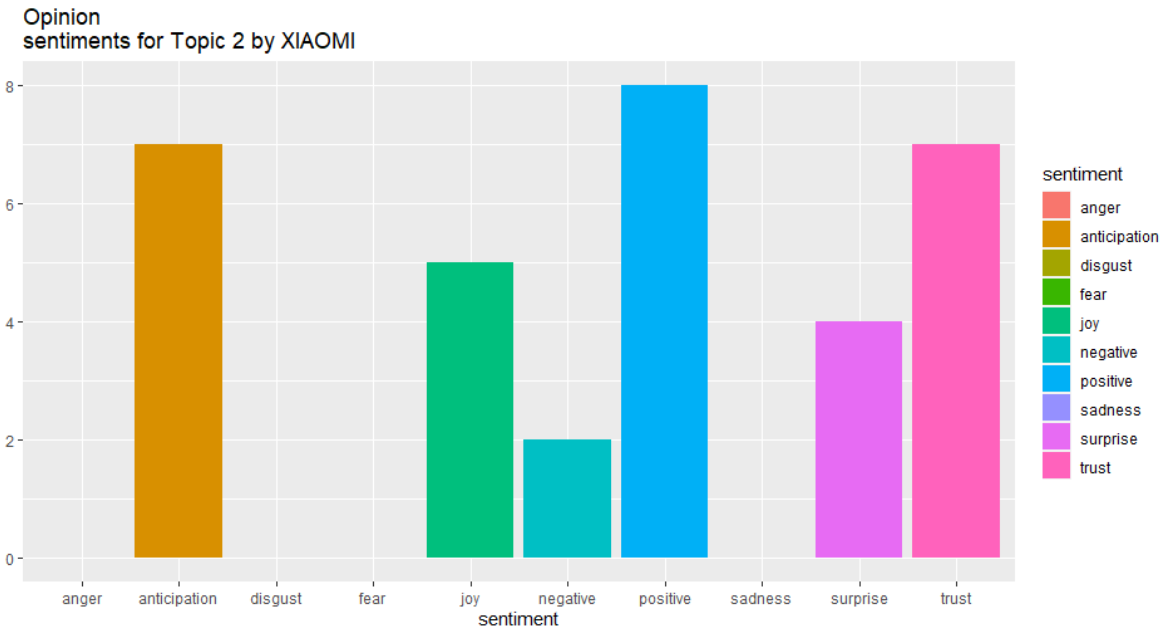
```
> neg1$Score
[1] -1
> length(neg1$Score)
[1] 1
```

Wyniki scoringu dla topic'u 2 kształtują się w atmosferze pozytywno-neutralnej. Na temat Wydajności pozytywnie wypowiedziało się 6 nabywców urządzenia, 4 osoby pozostały neutralnie nastawione oraz 1 wypowiedziała się w sposób negatywny.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych (po lewej stronie), neutralnych (pośrodku) i negatywnych (po prawej). Można z nich wnioskować, że klienci najbardziej cenili sobie dużą pamięć telefonu, wydajność procesora, oraz ogólne zadowolenie z produktu. Opinie neutralne wskazywały na to, czy telefon sprostał oczekiwaniom, podczas gdy negatywne głosy wspominały o problemach które występują podczas robienia zdjęć.

- **Opinion sentiments**

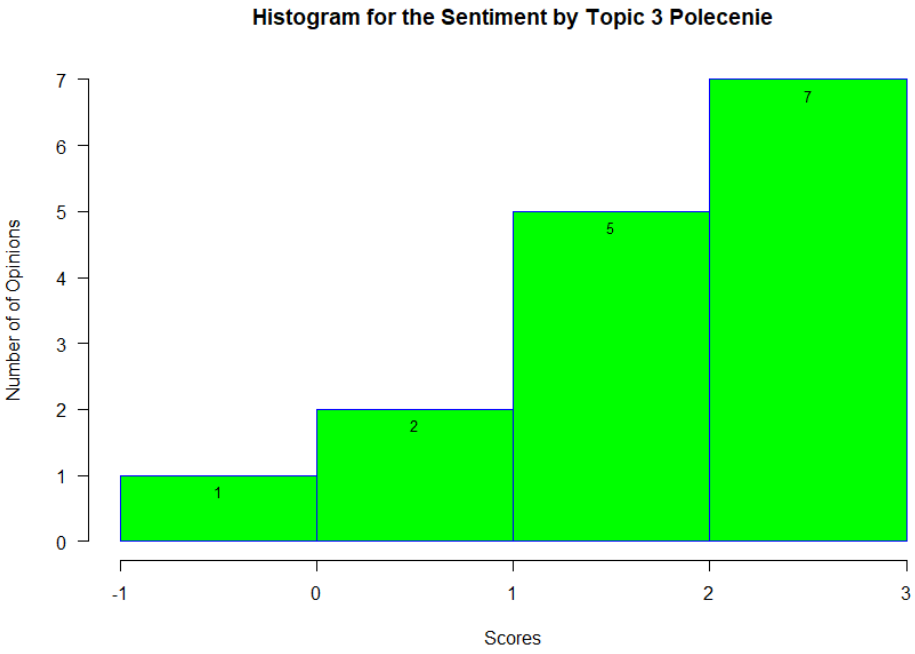


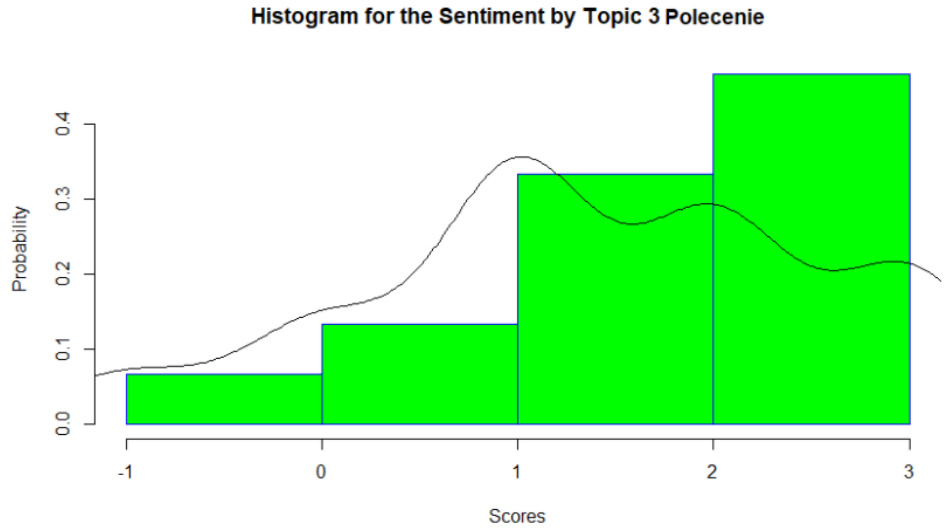
Sentyment w opiniach dla wydajności urządzenia osiągnął praktycznie sam wydźwięk pozytywny. Klienci byli zadowoleni z zakupu, jedynie tylko kilka reakcji przyjęło negatywne uczucia.

TOPIC 3 - Polecenie

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -1.0     1.0     1.0     1.4     2.0     3.0
> minn<-min(m1$Score)
> minn
[1] -1
> maxx<-max(m1$Score)
> maxx
[1] 3
```






```
> pos1$Score
 [1] 1 2 2 3 1 1 1 2 1 3 3 2
> length(pos1$Score)
[1] 12
```

```
> neu1$Score
[1] 0 0
> length(neu1$Score)
[1] 2
```

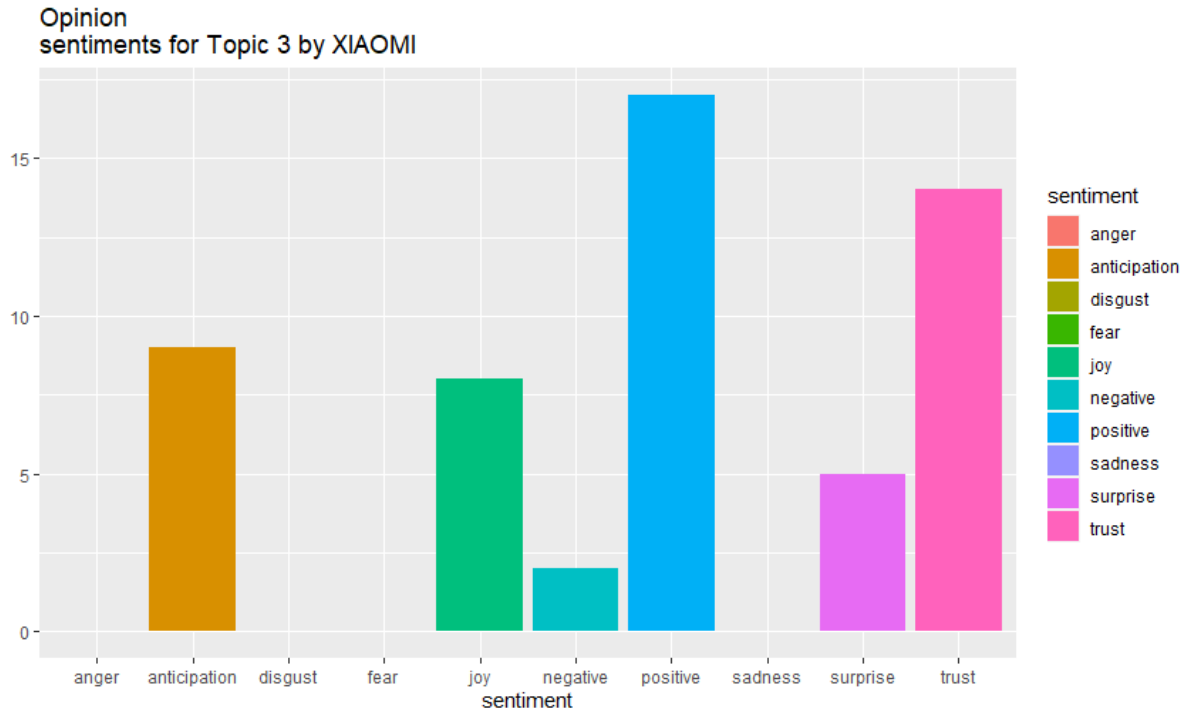
```
> neg1$Score
[1] -1
> length(neg1$Score)
[1] 1
```

Wyniki scoringu dla topic'u 3 kształtują się w atmosferze bardziej pozytywnej. Zauważa się, że pod względem polecenia oraz chwalenia modelu telefonu pojawiło się 12 opinii pozytywnych, 2 neutralne oraz 1 negatywna.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych (po lewej stronie), neutralnych (pośrodku) i negatywnych (po prawej). Można z nich wnioskować, że klienci najbardziej cenili sobie stosunek jakości do ceny, dzięki czemu chętnie rekomendowali omawiany produkt. Opinie neutralne wskazywały na szeroką gamę atutów telefonu, podczas gdy negatywne głosy mówiły o problemach systemowych smartphona.

- **Opinion sentiments**



Pod względem polecenia telefonu podobnie jak wcześniej sentyment także układa się w sferze pozytywnej. Wszystkie słupki odzwierciedlające pozytywne reakcje mają wysokie wskaźniki, co znaczy, że klienci chętnie polecali zakup telefonu Xiaomi Redmi Note 9 Pro 6.

5.4 HUAWEI - P40 Lite

DTM powstały dla HUAWEI P40 Lite dysponuje wymiarami w liczbie 50 opinii na 489 słów, gdzie rozproszenie macierzy (sparsity) wynosi 96%, a najdłuższe słowo ma długość 14-tu liter.

```
> dtm
<<DocumentTermMatrix (documents: 50, terms: 489)>>
Non-/sparse entries: 1010/23440
Sparsity           : 96%
Maximal term length: 14
Weighting          : term frequency (tf)
```

5.4.1 Statystyki opisowe dla opinii - HUAWEI

- Opinia o maksymalnej długości

```
> max_length<-max(doc_length)
> max_length
[1] 88
```

- Opinia o minimalnej długości

```
> min_length<-min(doc_length)
> min_length
[1] 1
```

- Średnia długość opinii w korpusie

```
> aver_length<-mean(rowSums(as.matrix(dtm)))
> aver_length
[1] 22.92
```

5.4.2 Preprocessing

- Przykładowa opinia przed preprocessingiem

```
> writeLines(as.character(docs[[11]]))
A very well-made phone. User-friendly interface and easy to learn. Cameras are a pure revelation. It's
 hard to find a better phone in this price range. I recommend.
```

- Opinia po przeprowadzeniu preprocessingu:

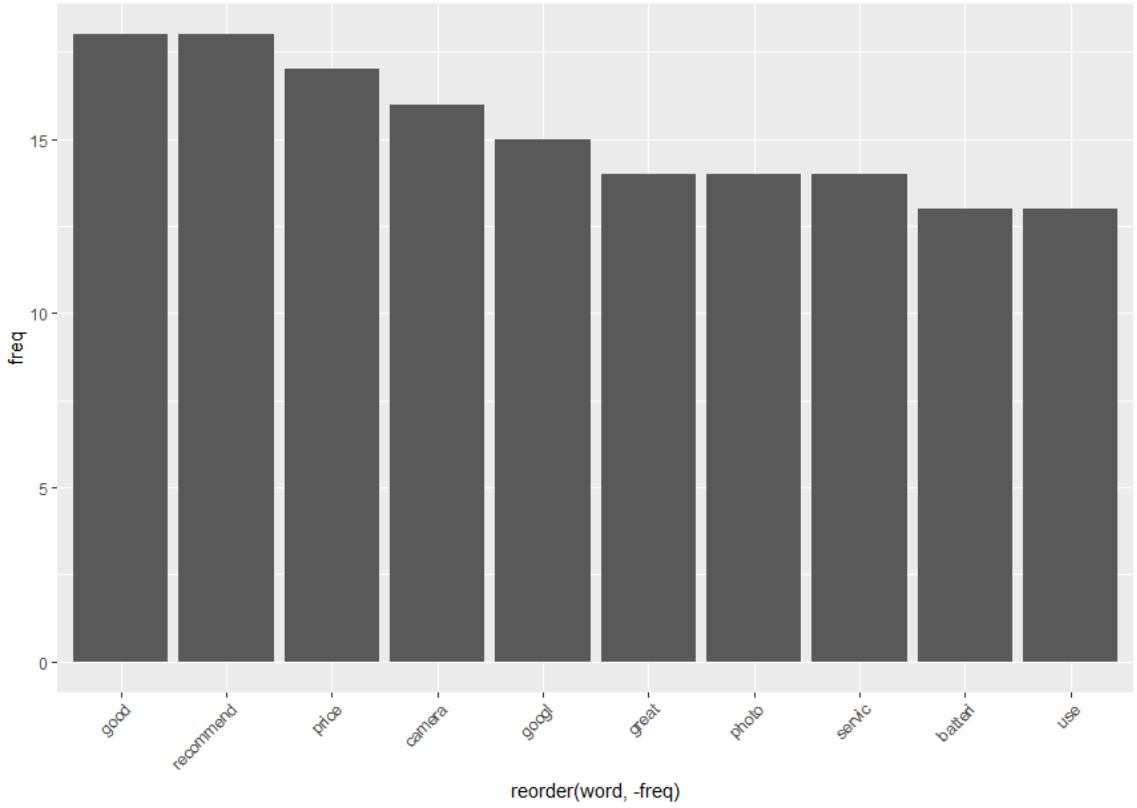
```
> writeLines(as.character(docs[[11]]))
wellmad userfriend interfac easi learn camera pure revel hard find better price rang recommend
```

5.4.3 DTM dla opinii po przeprowadzeniu procesu preprocessingu

```
> dtm
<<DocumentTermMatrix (documents: 50, terms: 113)>>
Non-/sparse entries: 471/5179
Sparsity           : 92%
Maximal term length: 10
Weighting          : term frequency (tf)
```

DTM powstały dla urządzenia Huawei po preprocessingu dysponuje wymiarami w liczbie 50 opinii na 113 słów, gdzie rozproszenie macierzy (sparsity) wynosi 92%, a najdłuższe słowo ma długość 10-tu liter.

5.4.4 Zipf



Najczęstszym słowem pojawiającym się w korpusie było “good”, następnie ”recommend” oraz “photo”. Częstotliwość występowania tych terminów pokazała pozytywny odbiór smartphona wśród nabywców oraz świadczy o tym, że klienci dzięki jakości jaką oferuje telefon i atrakcyjnej cenie polecają go do zakupu. Kolejne słowa potwierdziły tę tezę.

5.4.5 Wordcloud

Powyższe wordcloudy potwierdzają wynik osiągnięty w pierwszej metodzie graficznej. Biorąc pod uwagę przedstawione chmury słów, klienci w swoich opiniach wskazywali na pozytywny odbiór smartphona ze szczególnym uwzględnieniem funkcji aparatu, jakości baterii, serwisu oraz ceny, dzięki czemu chętnie polecali go i pisali o nim w superlatywach.

5.4.6 Topic Modelling

Przed podziałem recenzji na główne tematy zawsze sprawdzamy czy w korpusie znajdują się puste wiersze powstałe w rezultacie wstępnego przetwarzania opinii. Jak możemy odczytać z wyników otrzymanych w R nasz korpus zawiera dwa puste wiersze. Komentarze te zniknęły poprzez mechaniczne usunięcie słów, ponieważ znajdowało się w nich tylko jedno słowo o długości liter mniejszej niż 3 np. wyraz “ok”.

```
> mmm<-nrow(dtm[raw.sum==0,])
> mmm
[1] 2
```

Rozwiązaliśmy ten problem poprzez usunięcie pustych wierszy. Jak widać poniżej dtm zawiera teraz tylko 48 opinii, natomiast liczba słów pozostała bez zmian.

```
> dtm2
<<DocumentTermMatrix (documents: 48, terms: 113)>>
Non-/sparse entries: 471/4953
Sparsity           : 91%
Maximal term length: 10
Weighting          : term frequency (tf)
```

Podobnie jak dla poprzedniego urządzenia, za pomocą metody Topic Modelling powstały główne tematy zawierające najistotniejsze słowa występujące w korpusie opinii na temat urządzenia Huawei.

```
> ldaOut.terms
     Topic 1  Topic 2     Topic 3
[1,] "price"  "recommend" "use"
[2,] "camera" "googl"     "work"
[3,] "great"  "servic"    "lack"
[4,] "photo"  "good"      "play"
[5,] "day"    "applic"    "good"
```

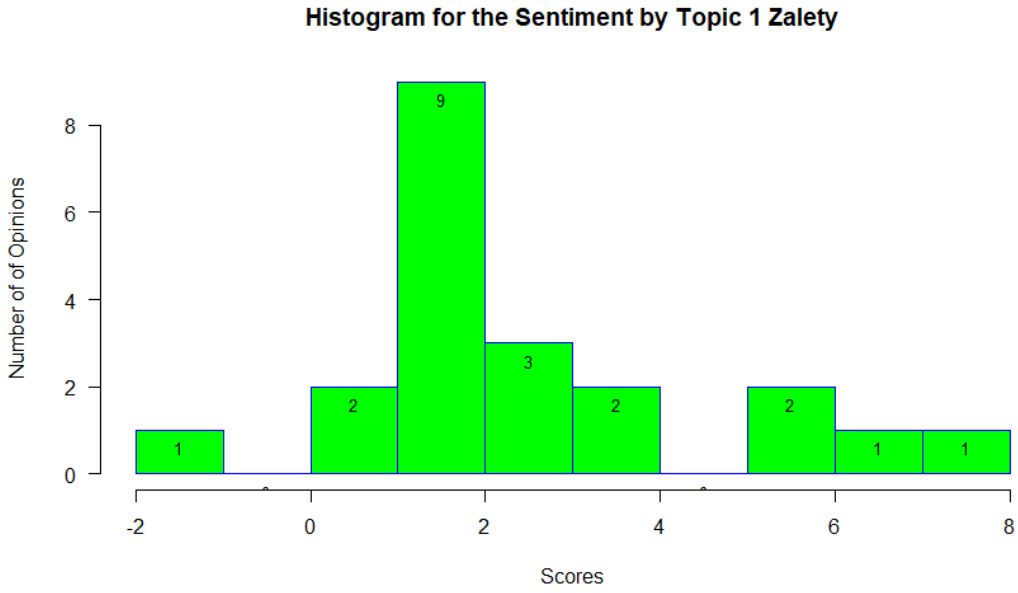
	TOPIC 1	TOPIC 2	TOPIC 3
NAZWA	Zalety	System	Wydajność
OPIS	Główne plusy smartphona	Serwis i aplikacje powiązane z oprogramowaniem	Wydajność telefonu oraz baterii podczas użytkowania
LICZBA OPINII	21	17	10
NUMER KOMENTARZA	11, 12, 14, 17, 18, 19, 20, 22, 24 ,25, 29, 32,34, 38, 39, 40, 41, 43, 44, 46, 47	2, 5, 9, 12, 15, 16, 23, 26, 27, 28, 30, 31, 33, 35, 37, 42, 45	1, 3, 4, 6, 7, 8, 10, 21, 36, 48

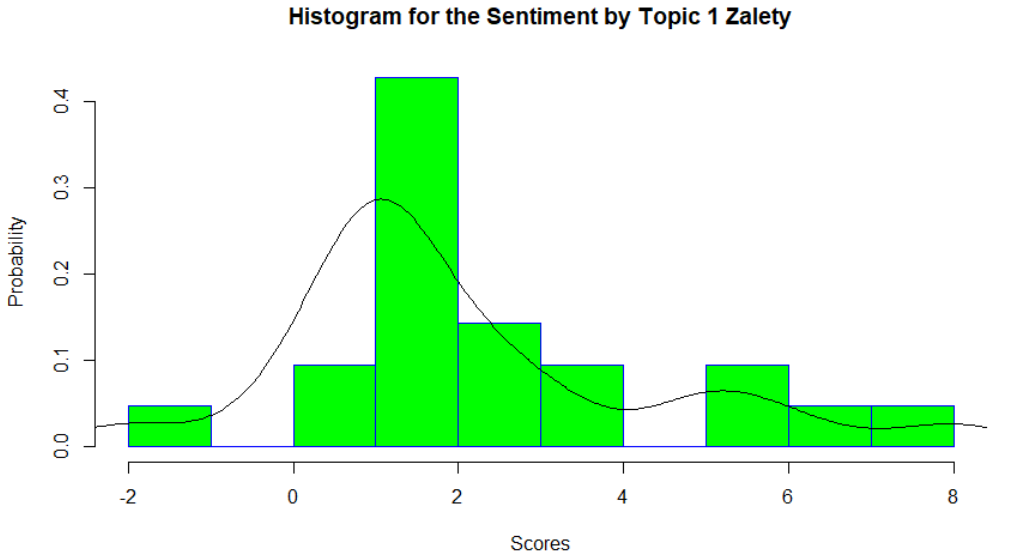
Najrzadziej pojawiającym się tematem wsród wszytskich opiniii okazała się wydajność telefonu, tylko 20.8% użytkowników wypowiedziało się w tej kwestii.

TOPIC 1 - Zalety

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -2.000   1.000   1.000   2.048   3.000   8.000
> minn<-min(m1$Score)
> minn
[1] -2
> maxx<-max(m1$Score)
> maxx
[1] 8
```






```
> pos1$score
 [1] 3 1 6 1 3 5 5 1 2 8 1 1 1 1 1 1 2 2
> length(pos1$score)
[1] 18
```

```
> neu1$Score
[1] 0 0
> length(neu1$Score)
[1] 2
```

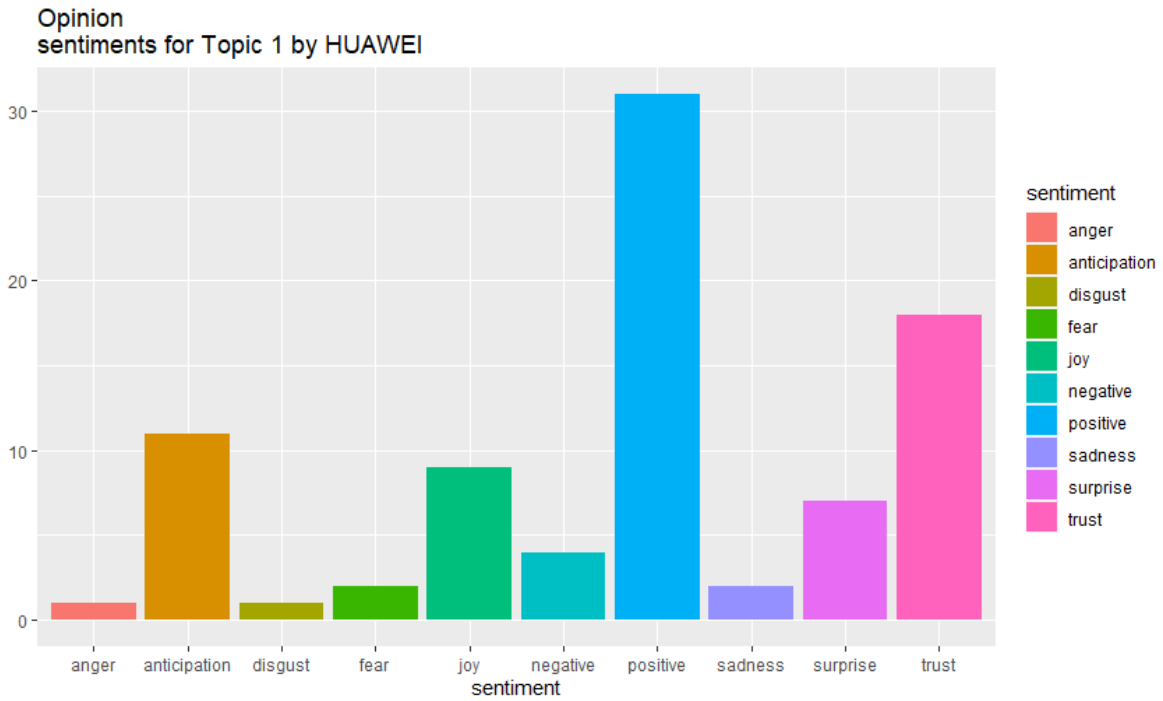
```
> neg1$Score
[1] -2
> length(neg1$Score)
[1] 1
```

Fundamentem po raz kolejny stały się opinie pozytywne. Tym razem wzięto pod uwagę zalety jakimi cechował się jeden z topowych urządzeń marki huawei. Znacznie częściej pojawiały się tu opinie w aurze pozytywnej, a ich liczba wyniosła aż 18 opinii. Kolejno pojawiły się 2 opinie neutralne oraz 1 negatywna.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych (po lewej stronie), neutralnych (pośrodku) i negatywnych (po prawej). Można z nich wnioskować, że klienci najbardziej cenili sobie aparat, ale również cenę i wydajność baterii, dzięki czemu chętnie rekomendowali produkt i pisali o nim w superlatywach. Opinie neutralne i negatywne wskazywały jednak na braki w dodatkowych funkcjach smartphona.

- **Opinion sentiments**

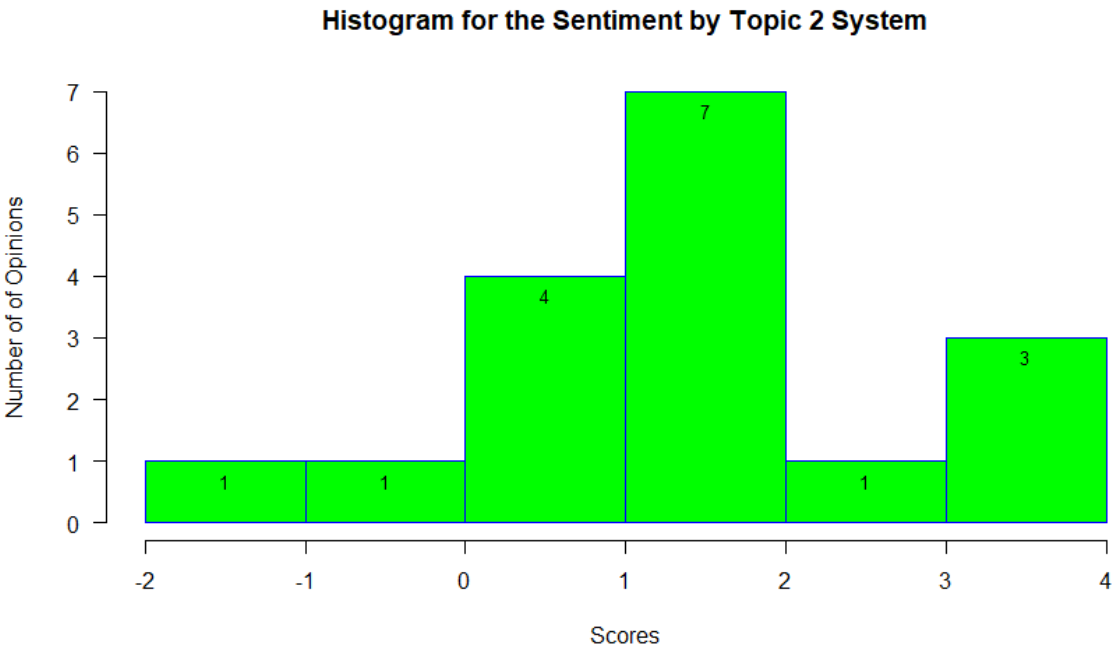


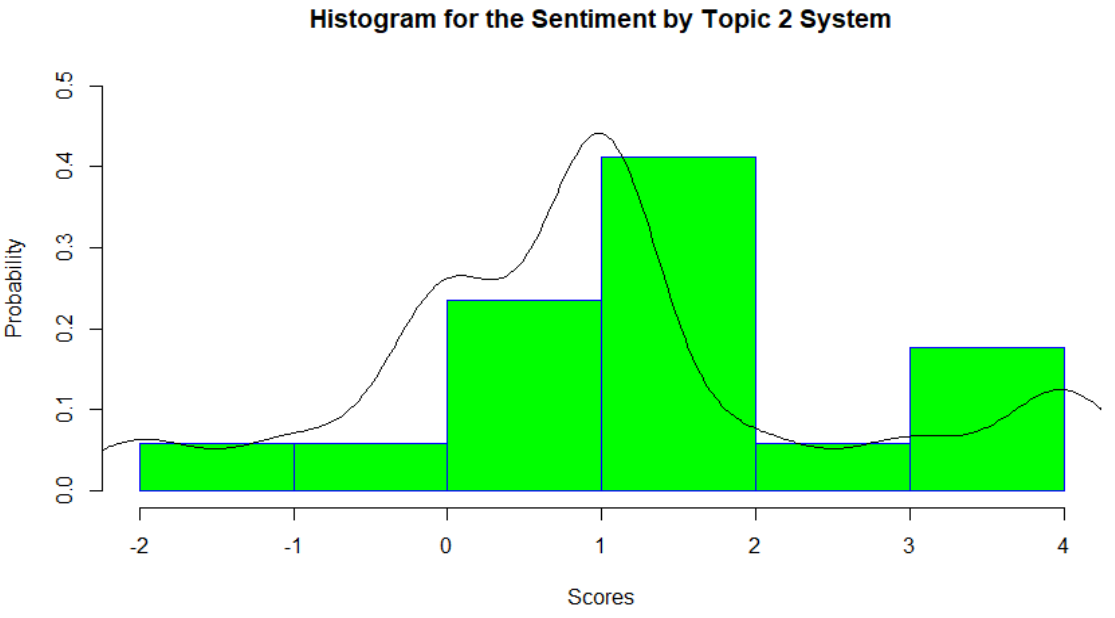
Sentyment dla topic'u 1 w Huawei wskazuje na pozytywne odczucia klientów odnośnie zakupu określonego modelu telefonu. Nabywcy byli zadowoleni, mile zaskoczeni, ufający marce oraz pozytywnie podchodzili do możliwości jakie prezentowało urządzenie. Odczucia negatywne w tym przypadku są na dość niskim poziomie co wskazuje, że klienci ogólnie nie byli rozczarowani i skupiali się na zaletach produktu.

TOPIC 2 - System

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     -2       0       1       1       1       4
> minn<-min(m1$Score)
> minn
[1] -2
> maxx<-max(m1$Score)
> maxx
[1] 4
```





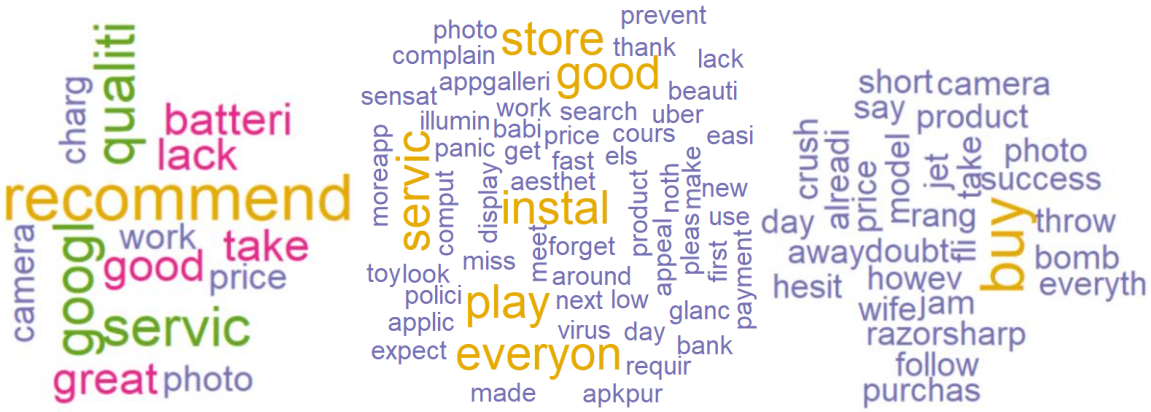
```
> pos1$Score
 [1] 1 1 1 4 4 1 2 1 1 3 1
> length(pos1$Score)
[1] 11
```

```
> neu1$Score
[1] 0 0 0 0
> length(neu1$Score)
[1] 4
```

```
> neg1$Score
[1] -2 -1
> length(neg1$Score)
[1] 2
```

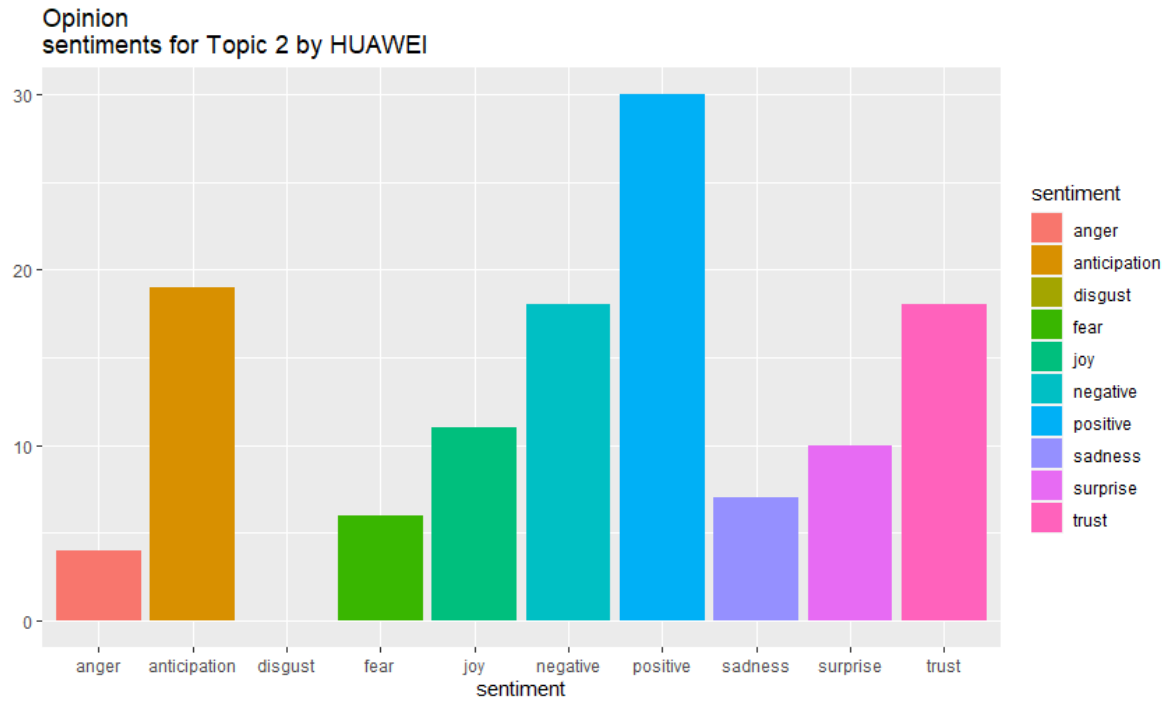
Wyniki scoringu dla topic'u 2 kształtują się w różnej atmosferze, lecz z widoczną większością opinii pozytywnych. Na temat Systemu pozytywnie wypowiedziało się 11 nabywców urządzenia, 4 osoby pozostały neutralnie nastawione oraz 2 wypowiedziała się w sposób negatywny.

- **Wordcloud**



Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych, neutralnych i negatywnych. Można z nich wnioskować, że klienci najbardziej cenili sobie dużą wydajność systemu na którym operuje smartphone, dzięki czemu chętnie go polecali. Opinie neutralne skupiały się na dyskusji odnośnie działania funkcji dodatkowych produktu, podczas gdy negatywne głosy wskazywały problemy występujące podczas korzystania z telefonu.

- **Opinion sentiments**

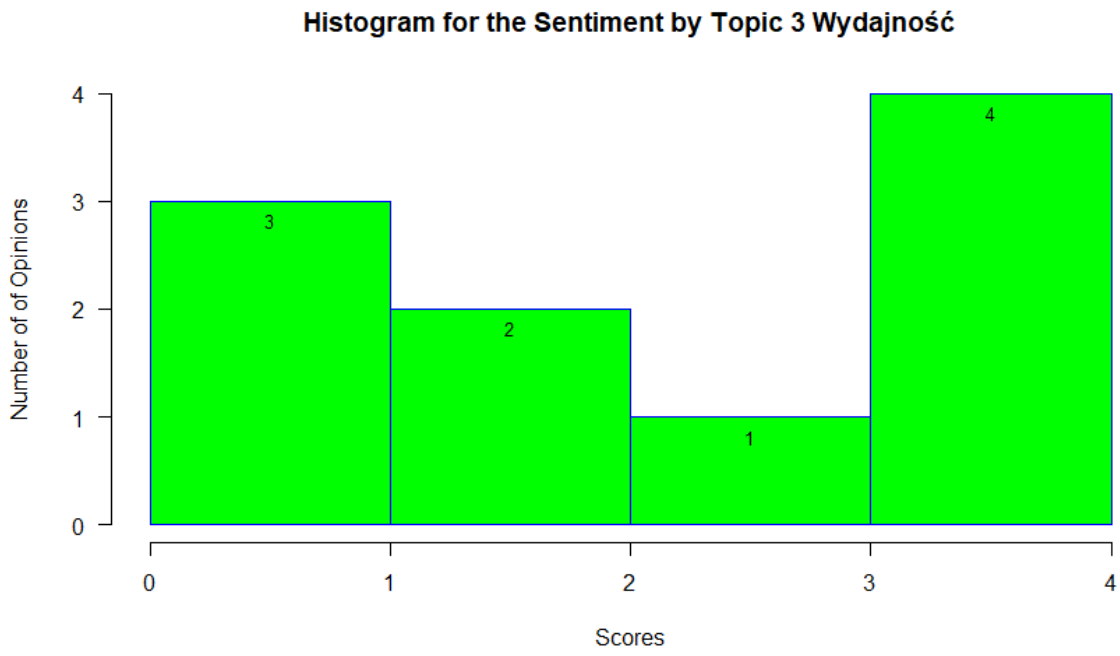


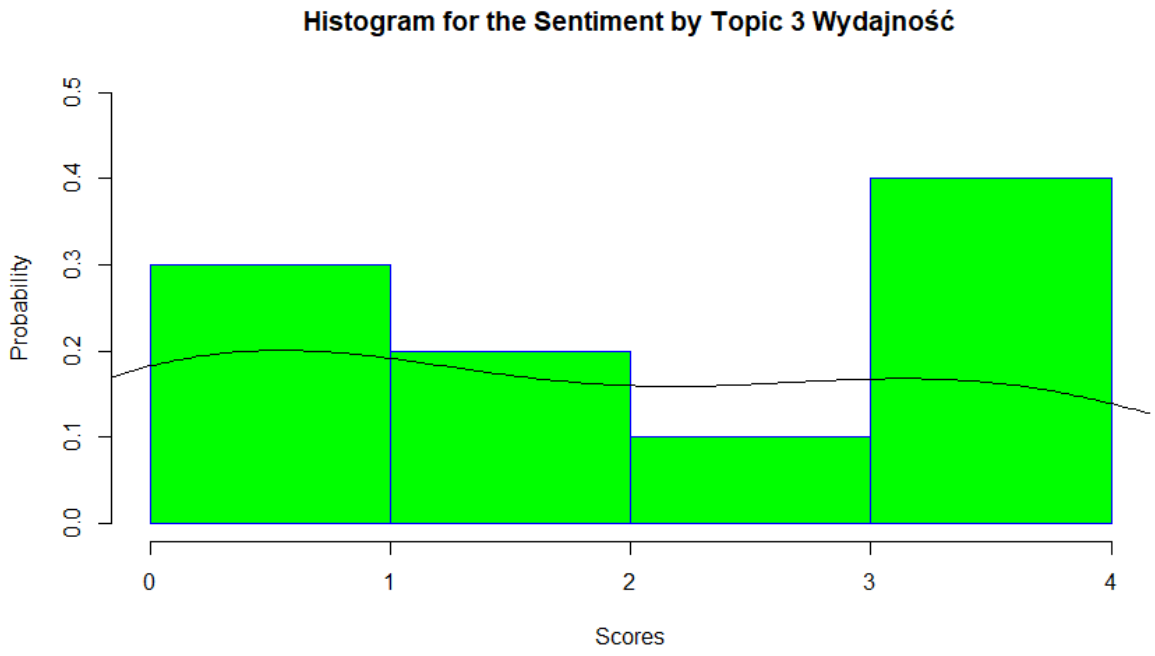
Sentyment dla topic'u 2 w Huawei wskazuje na pozytywne odczucia klientów odnośnie zakupu określonego modelu telefonu. Nabywcy byli zadowoleni, mile zaskoczeni oraz pozytywnie podchodzili do możliwości jakie reprezentowało urządzenie. Mimo to, odczucia negatywne w tym przypadku są na wysokim poziomie - w porównaniu do wcześniejszych przypadków, znacząco wzrosła złość oraz niepokój towarzyszący spadkowi poziomu zaufania.

TOPIC 3 - Wydajność

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.25    1.50    1.80    3.00    4.00
> minn<-min(m1$Score)
> minn
[1] 0
> maxx<-max(m1$Score)
> maxx
[1] 4
```






```
> pos1$Score
[1] 3 4 4 1 1 2 3
> length(pos1$Score)
[1] 7
```

```
> neu1$Score
[1] 0 0 0
> length(neu1$Score)
[1] 3
```

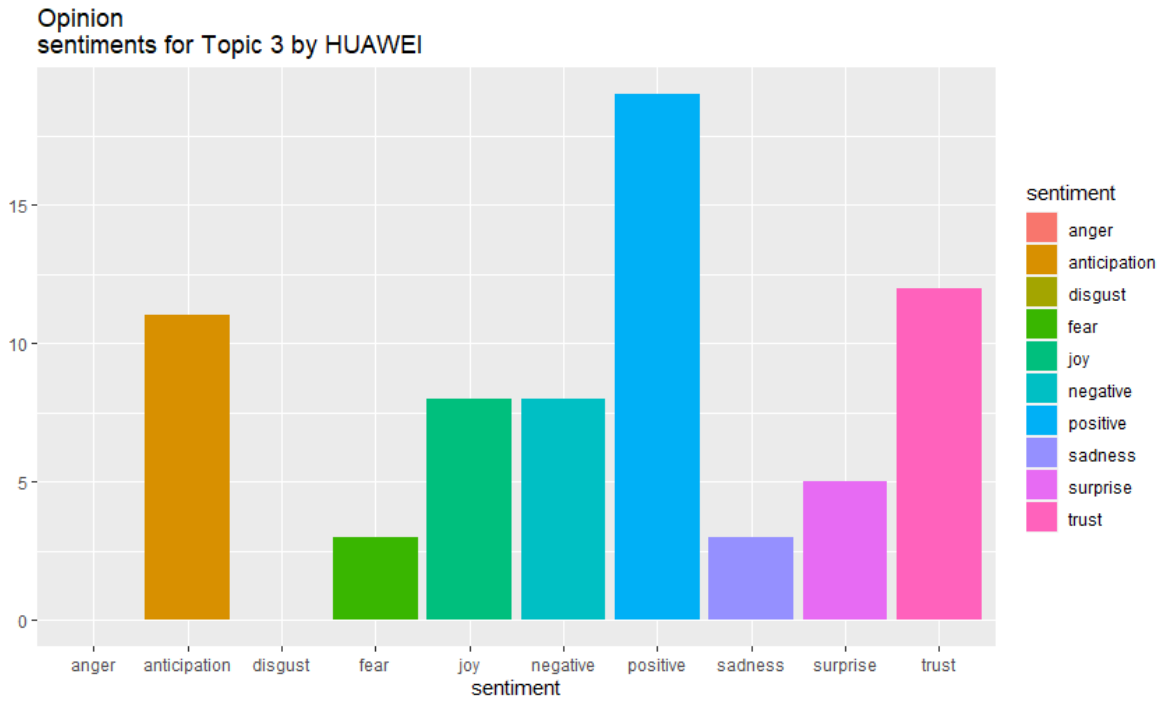
```
> neg1$Score
integer(0)
> length(neg1$Score)
[1] 0
```

Wyniki scoringu dla topic'u 3 kształtują się w atmosferze bardziej pozytywnej. Zauważa się, że pod względem wydajności model telefonu osiągnął 7 opinii pozytywnych oraz 3 neutralne.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych i neutralnych. Można z nich wnioskować, że klienci najbardziej cenili sobie wydajność produktu w codziennym użytkowaniu, dzięki czemu chętnie go rekomendowali. Opinie neutralne odnosiły się zaś do serwisu oferowanego przez firmę.

- **Opinion sentiments**



Sentyment dla Topic'u 3 w Huawei wskazuje na pozytywne odczucia klientów odnośnie zakupu określonego modelu telefonu. Nabywcy byli zadowoleni, mile zaskoczeni wydajnością telefonu oraz wykazywali duże zaufanie względem produktu. Odczucia negatywne które się pojawiły są na dość znaczącym poziomie, jednak nie przewyższającym superlatyw.

5.5 SAMSUNG - SM-A715 Galaxy A71

DTM powstały dla SAMSUNG SM-A715 Galaxy A71 dysponuje wymiarami w liczbie 50 opinii na 546 słów. Gdzie rozproszenie macierzy (sparsity) wynosi 96%, a najdłuższe słowo ma długość 14-tu liter.

```
> dtm
<<DocumentTermMatrix (documents: 50, terms: 546)>>
Non-/sparse entries: 1178/26122
Sparsity           : 96%
Maximal term length: 14
Weighting          : term frequency (tf)
```

5.5.1 Statystyki opisowe dla opinii - SAMSUNG

- Opinia o maksymalnej długości

```
> max_length<-max(doc_length)
> max_length
[1] 102
```

- Opinia o minimalnej długości

```
> min_length<-min(doc_length)
> min_length
[1] 2
```

- Średnia długość opinii w korpusie

```
> aver_length<-mean(rowSums(as.matrix(dtm)))
> aver_length
[1] 27.3
```

5.5.2 Preprocessing

- Przykładowa opinia przed preprocessingiem

```
> writeLines(as.character(docs[[1]]))
I heartily recommend the Samsung A715 phone is crazy. Before buying that phone, I wanted to buy a SLR b
ut with such a camera I do not need it. I would highly recommend.
```

- Opinia po przeprowadzeniu preprocessingu:

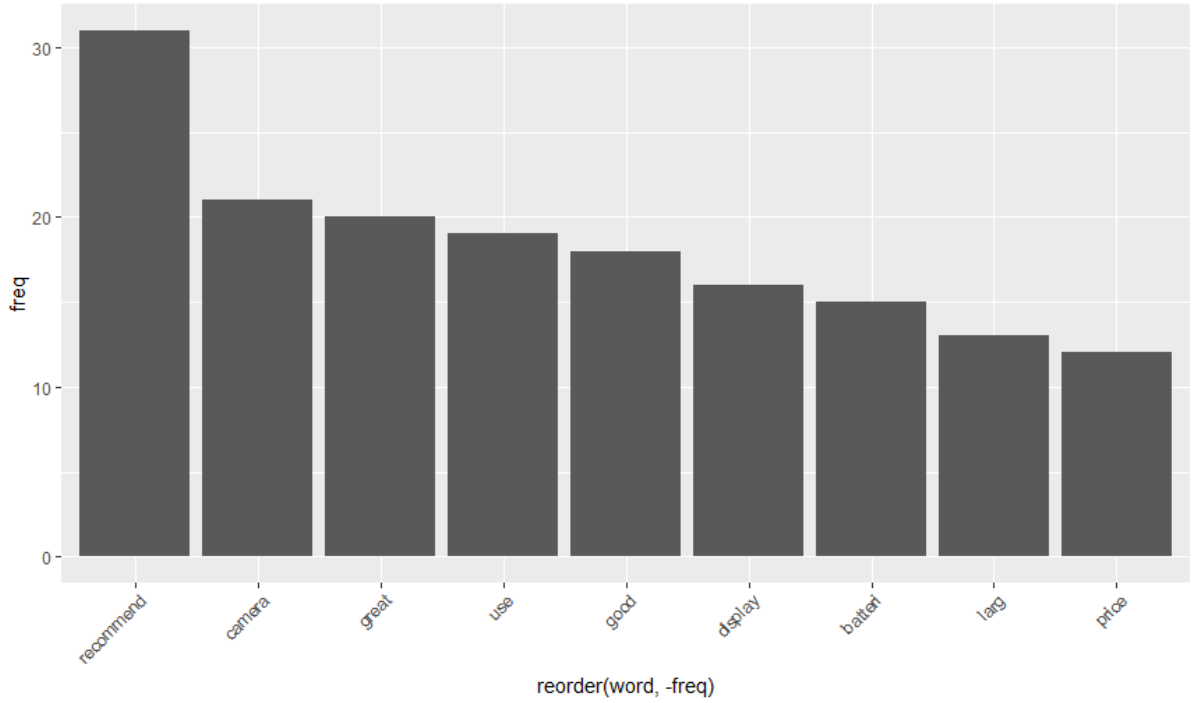
```
> writeLines(as.character(docs[[1]]))
heartili recommend crazi buy want buy slr camera need high recommend
```

5.5.3 DTM dla opinii po przeprowadzeniu procesu preprocessingu

```
> dtm
<<DocumentTermMatrix (documents: 50, terms: 141)>>
Non-/sparse entries: 586/6464
Sparsity           : 92%
Maximal term length: 11
Weighting          : term frequency (tf)
```

DTM powstały dla urządzenia SAMSUNG po preprocessingu dysponuje wymiarami w liczbie 50 opinii na 141 słów, gdzie rozproszenie macierzy (sparsity) wynosi 92%, a najdłuższe słowo ma długość 11-tu liter.

5.5.4 Zipf



Najczęstszym słowem pojawiającym się w korpusie było “recommend”, następnie ”camera” oraz “great”. Częstotliwość występowania tych terminów pokazała pozytywny odbiór smartphona wśród nabywców oraz świadczy o tym, że klienci dzięki jakości oraz funkcjom jakie oferuje telefon (z wyróżnieniem aparatu) polecają go do zakupu. Kolejne słowa potwierdziły tę tezę.

5.5.5 Wordcloud

Powyższe wordcloudy potwierdzają wynik osiągnięty w pierwszej metodzie graficznej. Biorąc pod uwagę przedstawione chmury słów, klienci w swoich opiniach wskazywali na jakość robionych zdjęć jak i ogólne zadowolenie z atrybutów telefonu, dzięki czemu chętnie polecali produkt i pisali o nim w superlatywach.

5.5.6 Topic Modelling

Podobnie jak dla poprzedniego urządzenia - za pomocą metody Topic Modelling powstały główne tematy zawierające najistotniejsze słowa występujące w korpusie opinii na temat marki SAMSUNG.

```
> ldaOut.terms
     Topic 1     Topic 2  Topic 3    Topic 4   Topic 5
[1,] "recommend" "great"  "use"      "batteri" "good"
[2,] "take"      "price"  "camera"   "larg"    "qualiti"
[3,] "buy"       "expert" "photo"    "display" "easi"
[4,] "display"   "nice"   "function" "model"   "realli"
[5,] "worth"     "beauti" "pleas"    "color"   "big"
```

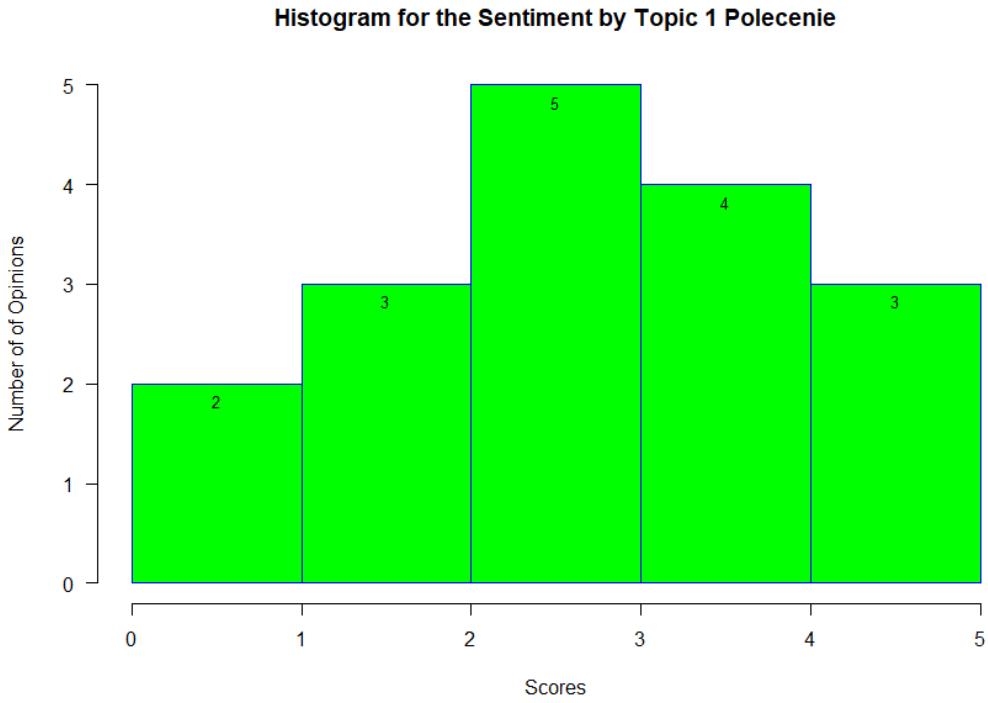
	TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4	TOPIC 5
NAZWA	POLECENIE	POZYTYWY	ZDJĘCIA	PARAMETRY	JAKOŚĆ
OPIS	Produkt jest warty zakupu	Zalety telefonu takie jak wygląd czy cena	Funkcje wbudowanego aparatu oraz jakość zdjęć	Dane techniczne i parametry modelu	Łatwość obsługi oraz jakość telefonu
LICZBA OPINII	17	9	6	11	7
NUMER KOMENTARZA	1, 2, 3, 9, 17, 19, 20, 26, 28, 29, 32, 33, 35, 38, 43, 47, 48	4, 5, 10, 11, 15, 18, 22, 31, 39	6, 36, 37, 40, 45, 46	14, 21, 23, 24, 25, 27, 41, 42, 44, 49, 50	8, 12, 13, 16, 30, 34

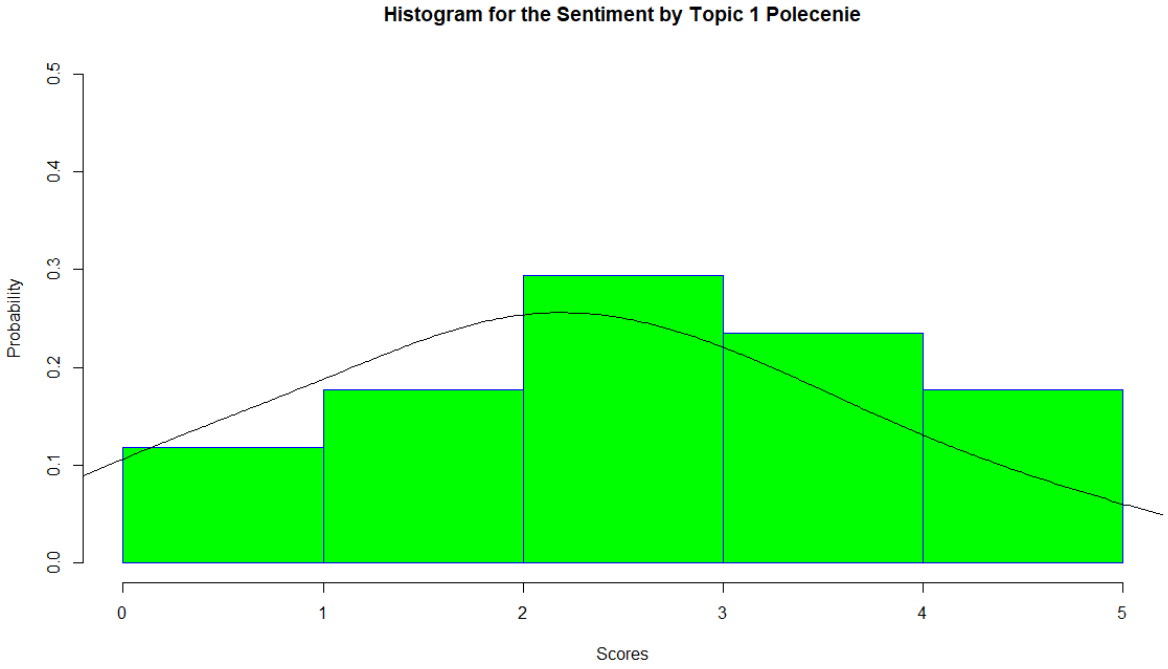
Po przeanalizowanych tematów możemy wywnioskować, że użytkownicy bardzo rzadko wypowiadali się na temat funkcji oraz obsługi telefonu, za to najwięcej komentarzy dotyczyło pozytywnej opinii i rekomendacji produktu.

TOPIC 1 - Polecenie

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   2.235   3.000   5.000
> minn<-min(m1$Score)
> minn
[1] 0
> maxx<-max(m1$Score)
> maxx
[1] 5
```






```
> pos1$Score
 [1] 2 4 1 3 3 1 2 5 1 4 3 2 3 2 2
> length(pos1$Score)
[1] 15
```

```
> neu1$Score
[1] 0 0
> length(neu1$Score)
[1] 2
```

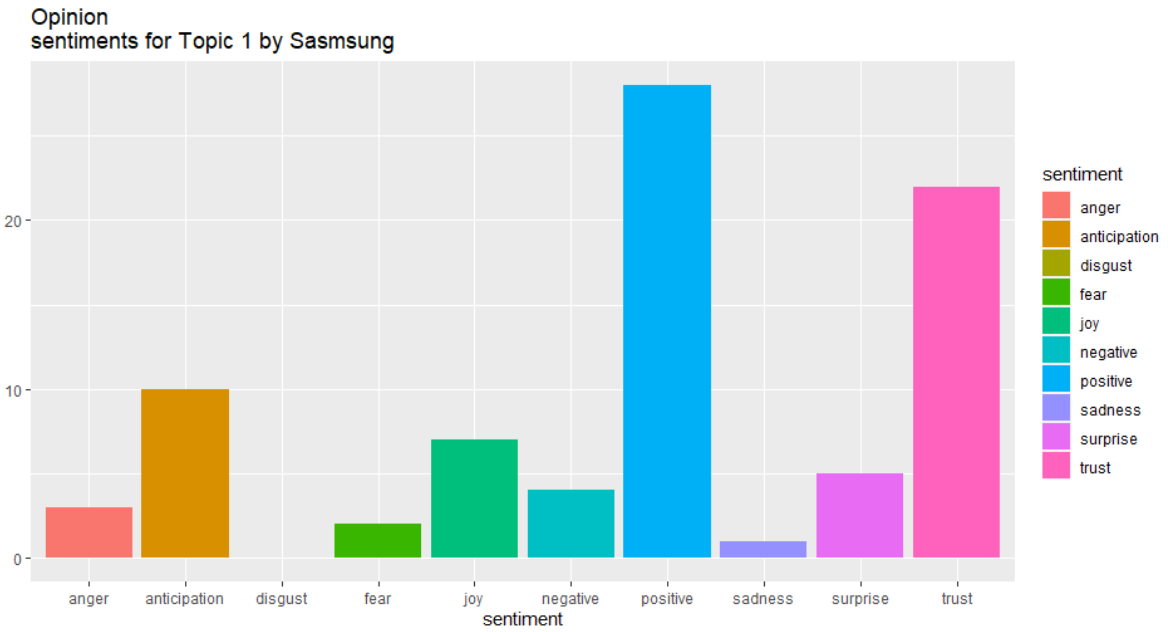
```
> neg1$Score
integer(0)
> length(neg1$Score)
[1] 0
```

W pierwszym topic'u dotyczącym telefonu marki Huawei pod względem polecenia go przez klientów innym osobom osiągnął 15 opinii pozytywnych oraz 2 neutralne. To świadczy, że Samsung - SM A715 Galaxy A71 osiągnął duży sukces komercyjny wśród jego nabywców.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych i neutralnych. Można z nich wnioskować, że klienci najbardziej cenili sobie pracę aparatu, wydajność baterii, ogólne użytkowanie smartphonu oraz stosunek ceny do jakości. Efektem było chętnie rekomendowanie go i pisanie o nim w superlatywach. Opinie neutralne również wskazują na pozytywne odczucia, jednakże z zaznaczeniem brakujących według użytkowników funkcji.

- **Opinion sentiments**

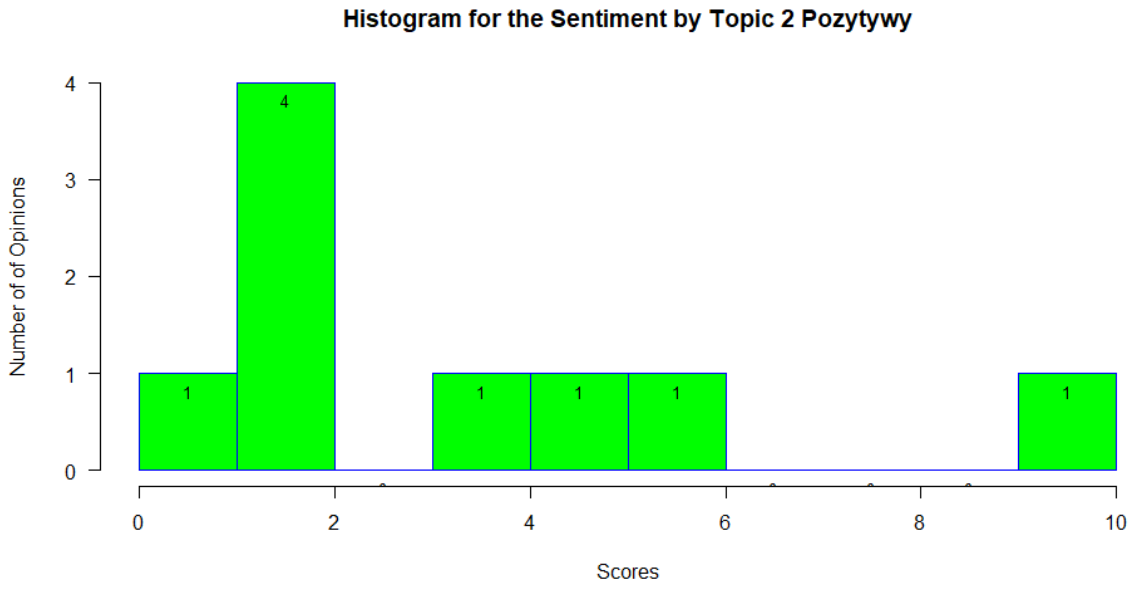


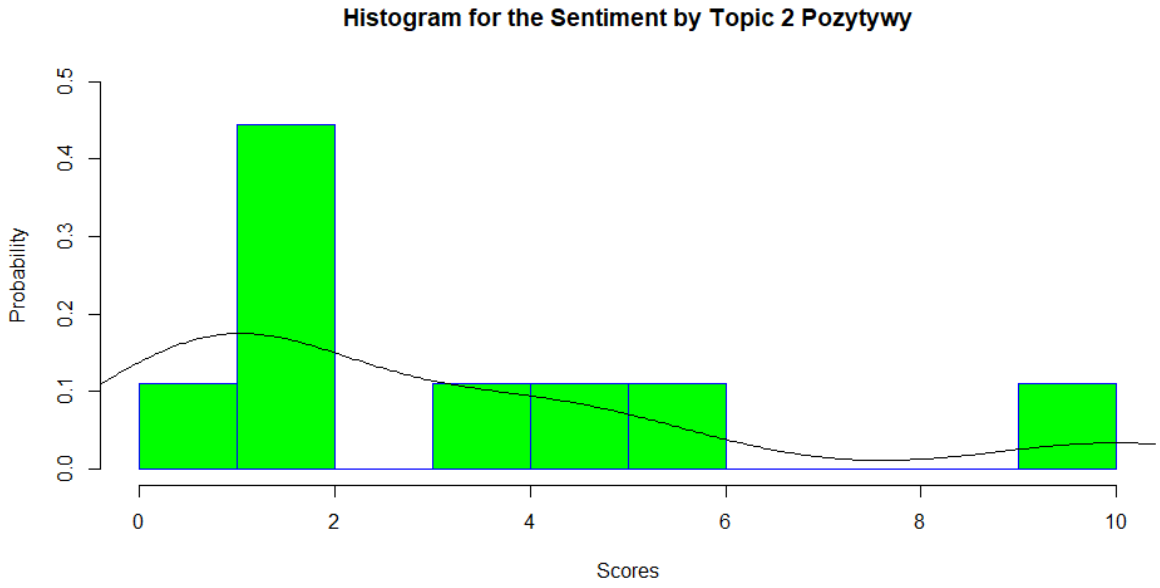
Sentyment dla topic'u 1 wskazał, że najwięcej opinii w tym temacie oscyluje na poziomie emocji silnie pozytywnych. Zauważalne jest to głównie przy dwóch słupkach histogramu, niebieskim – pozytyw oraz w kolorze magenty – zaufanie. Negatywne odczucia charakteryzują się niskimi wartościami.

TOPIC 2 - Pozytywy

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   1.000   2.889   4.000  10.000
> minn<-min(m1$Score)
> minn
[1] 0
> maxx<-max(m1$Score)
> maxx
[1] 10
```





```
> pos1$Score
[1]  5  1  3  1  1 10  1  4
> length(pos1$Score)
[1] 8
```

```
> neu1$Score
[1] 0
> length(neu1$Score)
[1] 1
```

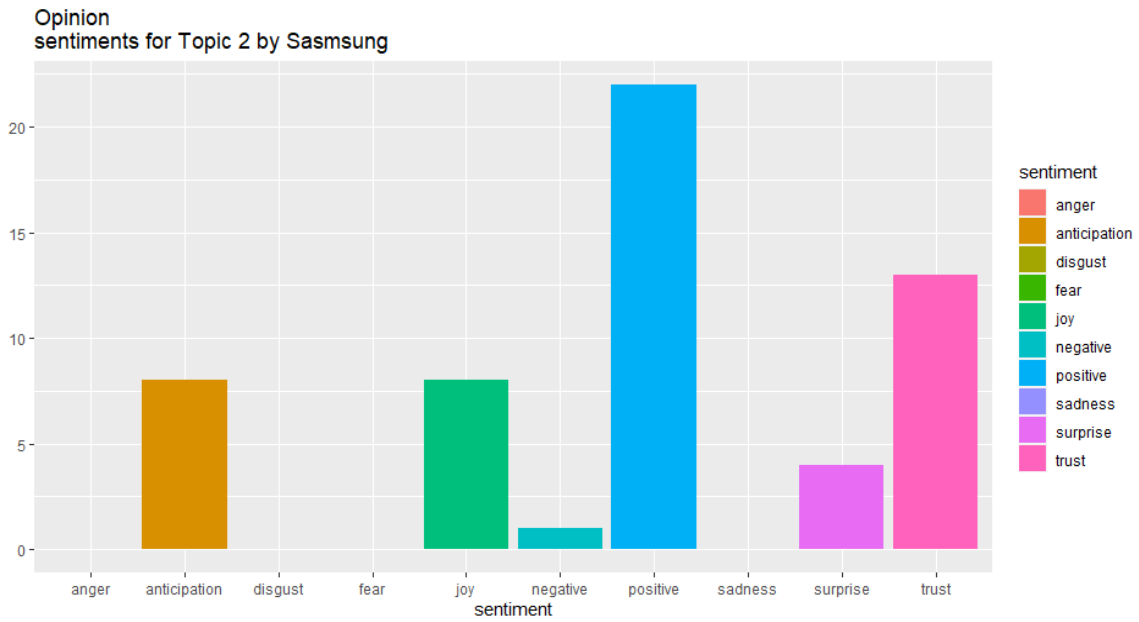
```
> neg1$Score
integer(0)
> length(neg1$Score)
[1] 0
```

Topic 2 oparty na zaletach telefonu takich jak wygląd oraz ocena, w 8 na 9 opinii uzyskał reakcje pozytywne oraz 1 neutralną. Smartphone według klientów był bardzo dobrze zaprojektowany wizualnie oraz jego cena była bardzo dobra w stosunku do tego czym reprezentował się telefon.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych i neutralnych. Można z nich wnioskować, że klienci najbardziej cenili sobie ogólne dobre wrażenia związane z korzystaniem z telefonu, na podstawie których polecali go znajomym. Opinie neutralne wskazały zaś na zarówno plusy jak i minusy produktu, zwracając uwagę na aspekty takie jak cena, bateria, oraz czytnik odcisków palców.

- **Opinion sentiments**

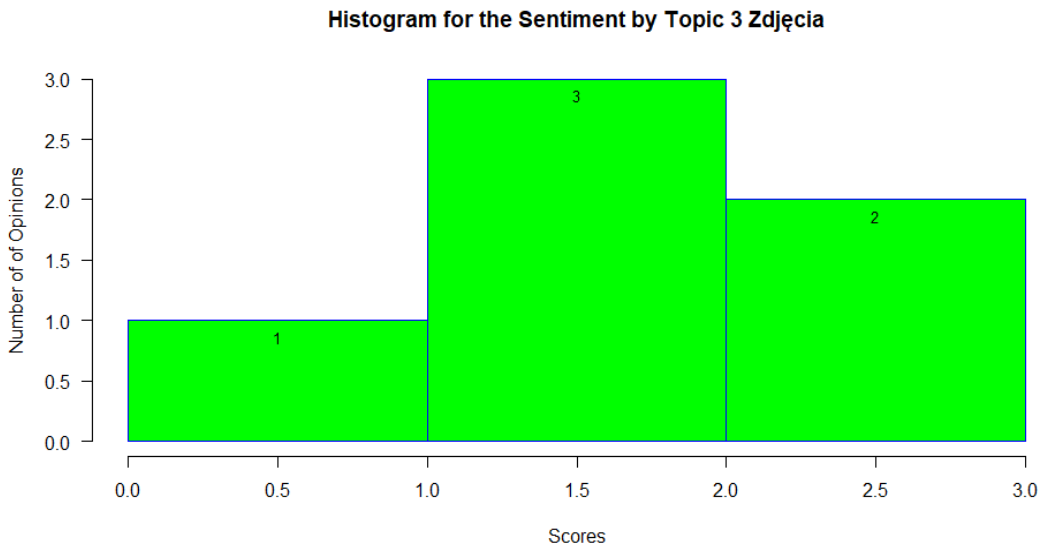


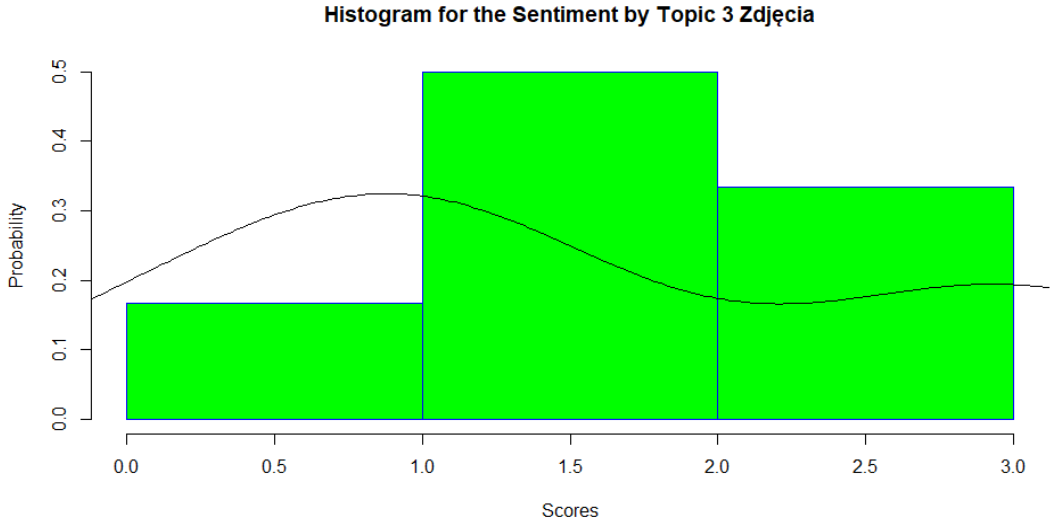
Sentyment dla topic'u 2 wskazał, że najwięcej opinii w tym temacie oscyluje na poziomie emocji silnie pozytywnych. Zauważalne jest to głównie przy dwóch słupkach histogramu, niebieskim – pozytyw, oraz w kolorze magenty – zaufanie. Negatywne odczucia praktycznie i znacząco nie występują.

TOPIC 3 - Zdjęcia

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     1.0     1.0     1.5     2.5     3.0
> minn<-min(m1$Score)
> minn
[1] 0
> maxx<-max(m1$Score)
> maxx
[1] 3
```





```
> pos1$Score
[1] 1 3 1 1 3
> length(pos1$Score)
[1] 5
```

```
> neu1$Score
[1] 0
> length(neu1$Score)
[1] 1
```

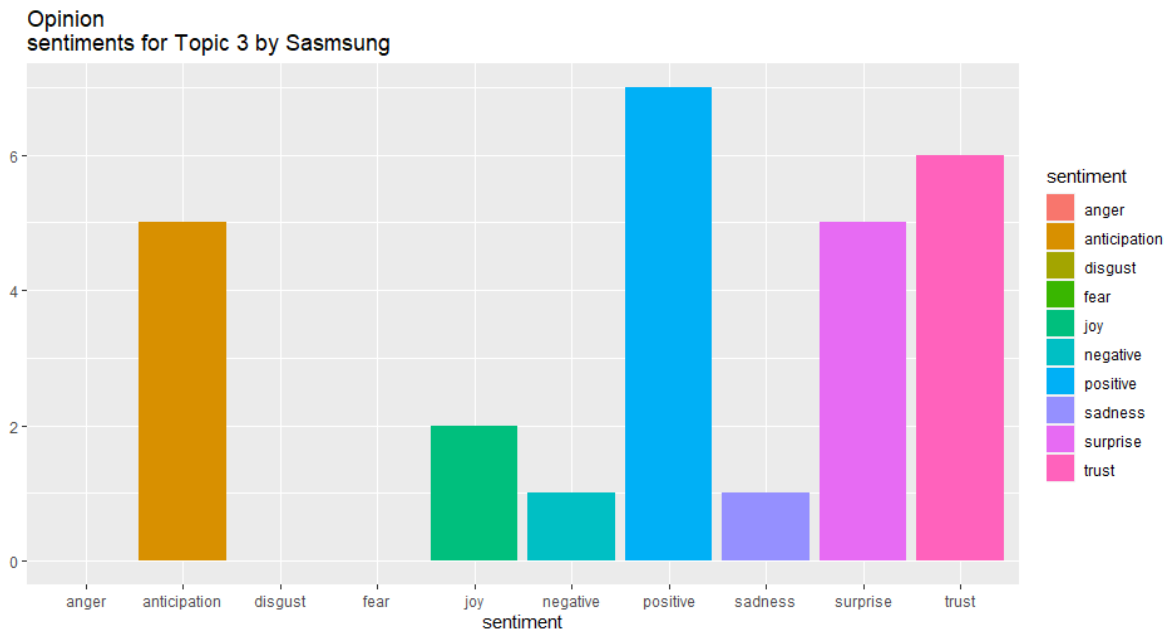
```
> neg1$Score
integer(0)
> length(neg1$Score)
[1] 0
```

Topic 3 oraz opinie, z których jest zbudowany wskazują na zadowolenie klientów pod względem wydajności jaką osiąga model marki Samsung. Temat zbudowany został z 6 opinii, gdzie 5 to reakcje pozytywne oraz 1 neutralna.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych i neutralnych. Można z nich wnioskować, że klienci najbardziej cenili sobie wysoką jakość kamery i wykonywanych zdjęć. W opiniach neutralnych użytkownicy polecali produkt, jednak wspomnali również o problemach wynikających z braku atutu jego wodoodporności.

- **Opinion sentiments**

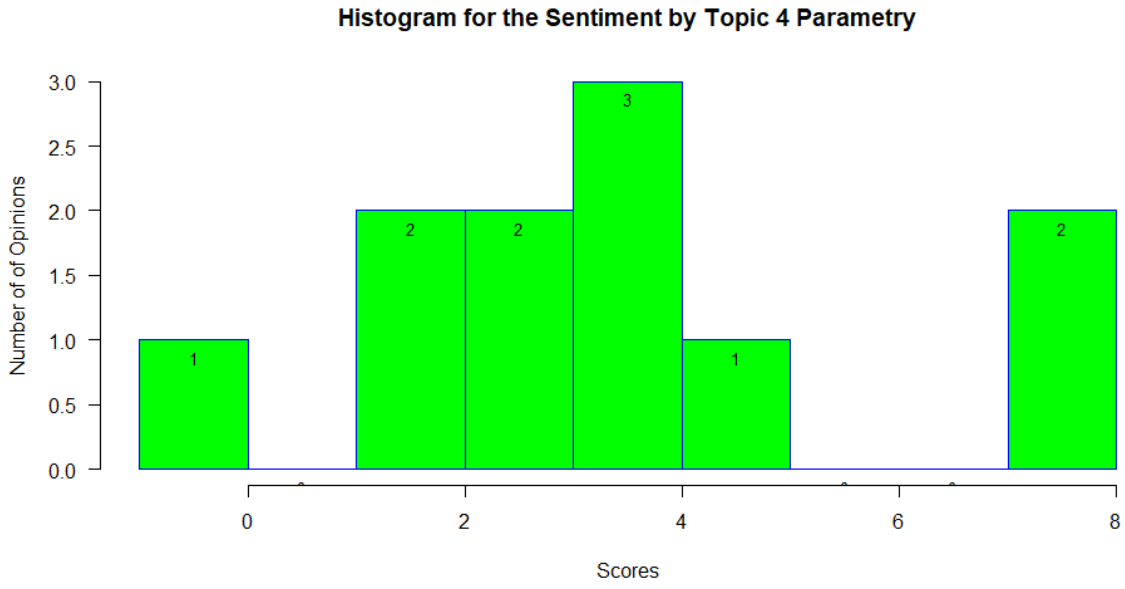


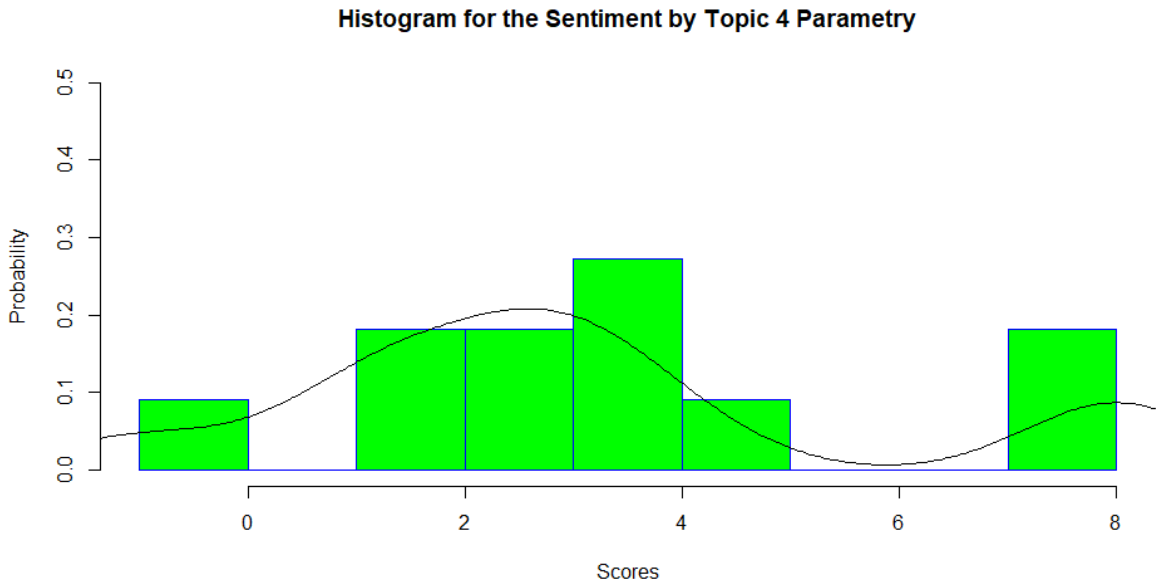
Sentyment dla topic'u 3 wskazał, że najwięcej opinii w tym temacie oscyluje na poziomie emocji silnie pozytywnych. Zauważalne jest to głównie przy słupkach histogramu: niebieskim – pozytyw oraz w kolorze magenty – zaufanie (duże wartości można zaobserwować również dla pomarańczowego - oczekiwanie, oraz różowego - zaskoczenie). Negatywne odczucia charakteryzują się niskimi lub zerowymi wartościami.

TOPIC 4 - Parametry

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.000   1.500   3.000   3.091   3.500   8.000
> minn<-min(m1$Score)
> minn
[1] -1
> maxx<-max(m1$Score)
> maxx
[1] 8
```






```
> pos1$Score
 [1] 2 1 1 3 3 3 8 2 4 8
> length(pos1$Score)
[1] 10
```

```
> neu1$Score
integer(0)
> length(neu1$Score)
[1] 0
```

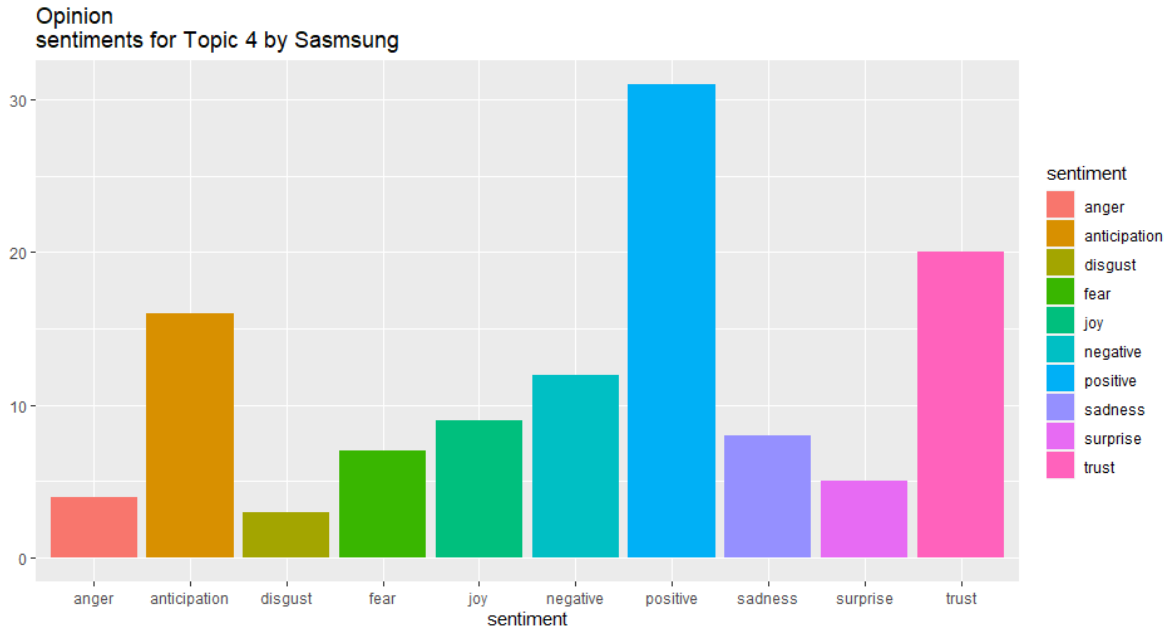
```
> neg1$Score
[1] -1
> length(neg1$Score)
[1] 1
```

W topic'u 4 opinie uformowały temat odnoszący się do parametrów telefonu oraz danych technicznych. W większości klienci pozytywnie odnosili się pod tym względem, jednak pojawiła się jedna opinia negatywna.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych i neutralnych. Można z nich wnioskować, że klienci najbardziej cenili sobie wysoką wydajność oraz możliwości smartphonu, z uwzględnieniem szerokiego wachlarza funkcji. Opinie neutralne zaś wskazały na istnienie drobnych problemów występujących podczas korzystania z urządzenia.

- **Opinion sentiments**

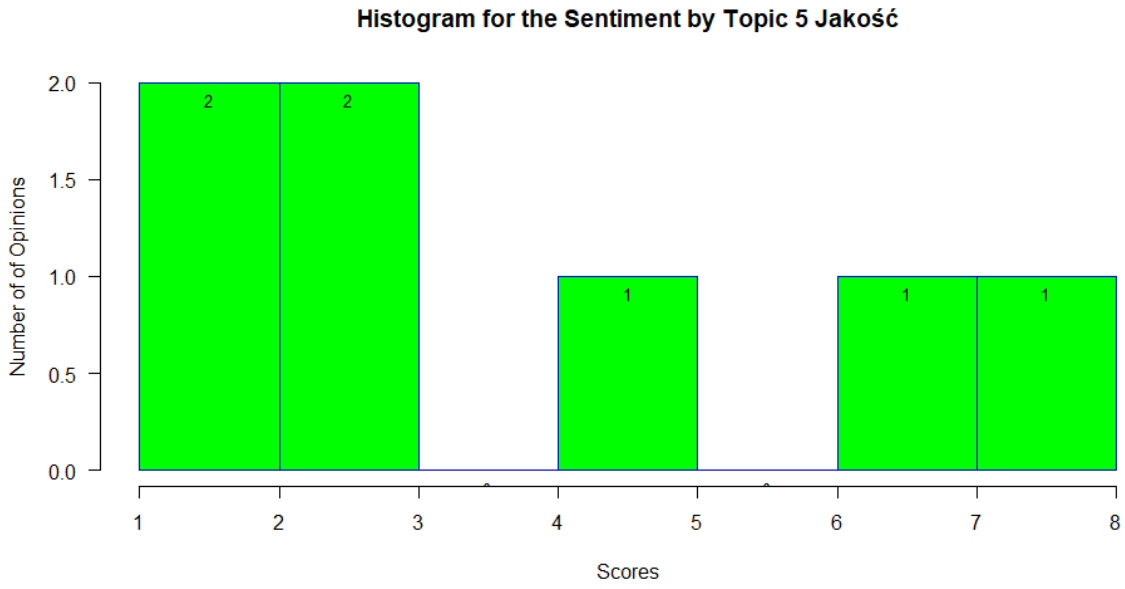


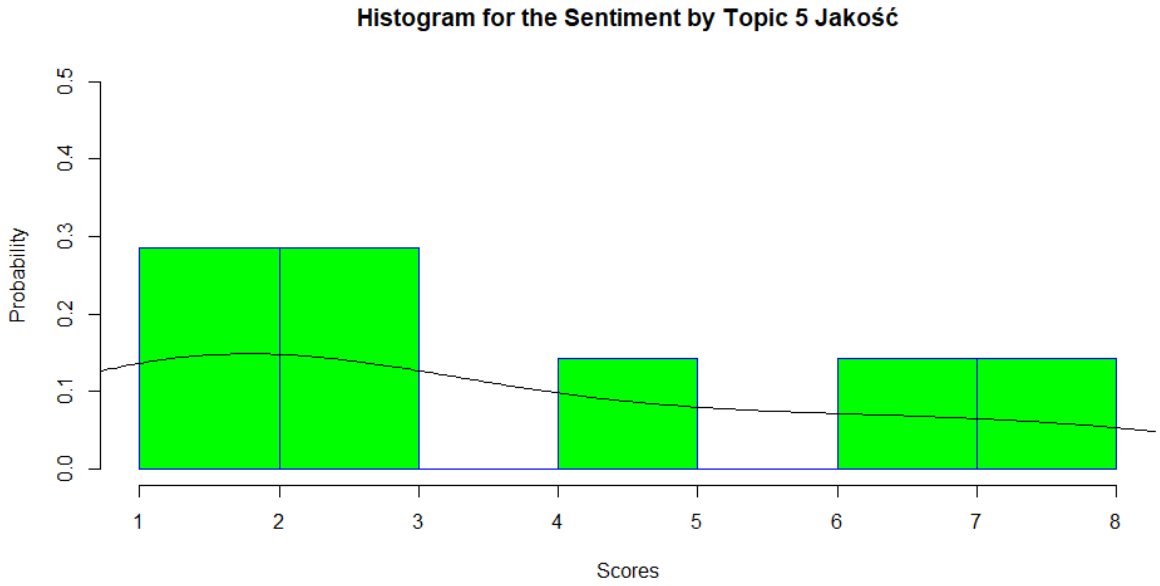
Sentyment dla topic'u 4 wskazał, że najwięcej opinii w tym temacie oscyluje na poziomie emocji pozytywnych. Zauważalne jest to głównie przy trzech słupkach histogramu, odnoszących się do pozytywów . Można jednak zauważyć, że w przeciwieństwie do poprzednich przypadków - pojawia się tu więcej negatywnych odczuć, mimo iż charakteryzują się one niskimi wartościami.

TOPIC 5 - Jakość

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.500   2.000   3.429   5.000   8.000
> minn<-min(m1$Score)
> minn
[1] 1
> maxx<-max(m1$Score)
> maxx
[1] 8
```






```
> pos1$Score
[1] 2 6 4 1 8 1 2
> length(pos1$Score)
[1] 7
```

```
> neu1$Score
integer(0)
> length(neu1$Score)
[1] 0
```

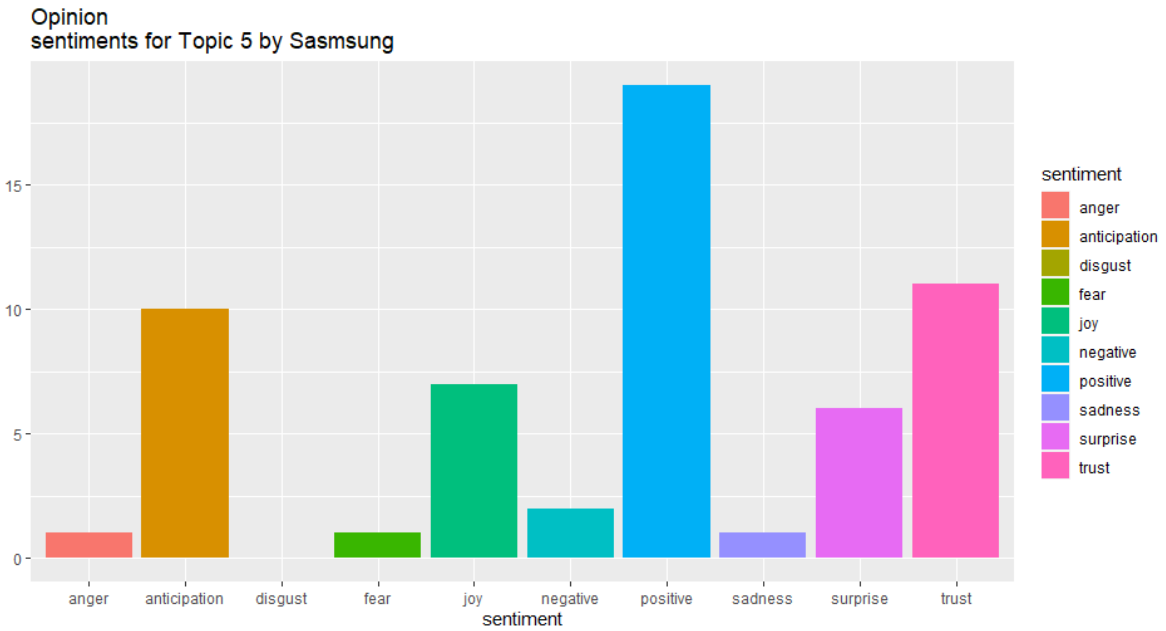
```
> neg1$Score
integer(0)
> length(neg1$Score)
[1] 0
```

Topic 5 to głównie tematyka związana z jakością oraz łatwością obsługi telefonu. Pod tym względem nabywcy tego modelu byli jednogłośni w chwaleniu urządzenia oraz przedstawieniu go w opiniach w samych pozytywach.

- **Wordcloud**

Powyższy wordcloud graficznie przedstawia częstotliwości występowania słów w opiniach pozytywnych. Można z niego wnioskować, że klienci najbardziej cenili sobie wysoką jakość oraz możliwości smartphonu (ze szczególnym wyróżnieniem wyświetlanego obrazu i kamery), dzięki czemu chętnie go rekomendowali.

- **Opinion sentiments**



Sentyment dla topic'u 5 wskazuję, że najwięcej opinii ma wysoki poziom odczuć pozytywnych. Wskaźniki takie jak oczekiwania, zadowolenie, pozytyw oraz zaufanie Negatywne odczucia charakteryzują się niskimi wartościami.

5.6 MOTOROLA - Moto G8 Power

DTM powstały dla MOTOROLA Moto G8 Power dysponuje wymiarami w liczbie 50 opinii na 530 słów, gdzie rozproszenie macierzy (sparsity) wynosi 96%, a najdłuższe słowo ma długość 16-tu liter.

```
> dtm
<<DocumentTermMatrix (documents: 50, terms: 530)>>
Non-/sparse entries: 978/25522
Sparsity           : 96%
Maximal term length: 16
Weighting          : term frequency (tf)
```

5.6.1 Statystyki opisowe dla opinii - MOTOROLA

- Opinia o maksymalnej długości

```
> max_length<-max(doc_length)
> max_length
[1] 89
```

- Opinia o minimalnej długości

```
> min_length<-min(doc_length)
> min_length
[1] 2
```

- Średnia długość opinii w korpusie

```
> aver_length<-mean(rowSums(as.matrix(dtm)))
> aver_length
[1] 21.94
```

5.6.2 Preprocessing

- Przykładowa opinia przed preprocessingiem

```
> writeLines(as.character(docs[[3]]))
Excellent phone to my measure :-) I don't see the downsides.
```

- Opinia po przeprowadzeniu preprocessingu:

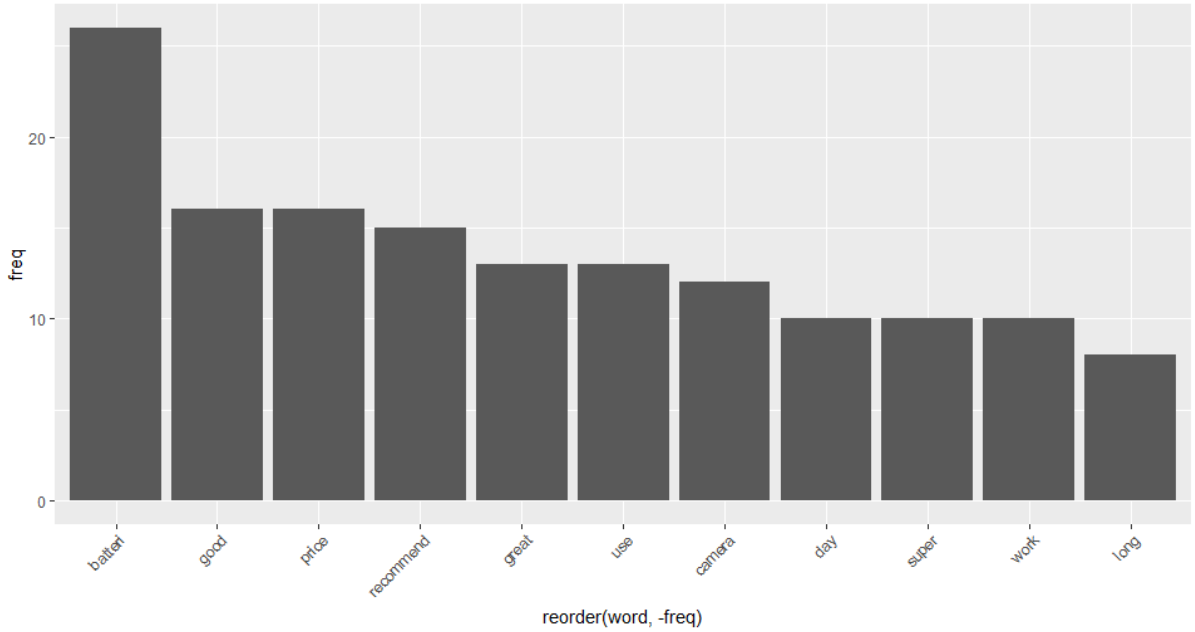
```
> writeLines(as.character(docs[[3]]))
excel measur dont see downsid
```

5.6.3 DTM dla opinii po przeprowadzeniu procesu preprocessingu

```
<<DocumentTermMatrix (documents: 50, terms: 115)>>
Non-/sparse entries: 440/5310
Sparsity           : 92%
Maximal term length: 9
Weighting          : term frequency (tf)
```

DTM powstały dla urządzenia Motorola po preprocessingu dysponuje wymiarami w liczbie 50 opinii na 115 słów, gdzie rozproszenie macierzy (sparsity) wynosi 92%, a najdłuższe słowo ma długość 9-ciu liter.

5.6.4 Zipf



Najczęstszym słowem pojawiającym się w korpusie było “batteri”, następnie ”good” oraz “price”. Częstotliwość występowania tych terminów pokazała pozytywny odbiór smartphona wśród nabywców oraz świadczy o tym, że klienci dzięki wysokiej klasy baterii oraz stosunku ceny do jakości polecają zakup produktu. Kolejne słowa potwierdziły tę tezę.

5.6.5 Wordcloud

Powyższe wordcloudy potwierdzają wynik osiągnięty w pierwszej metodzie graficznej. Biorąc pod uwagę przedstawione chmury słów, klienci w swoich opiniach wskazywali na wysokiej klasy baterię oraz stosunek ceny do jakości, jak również zachwalali działanie aparatu, a to przełożyło się na polecanie produktu i pisanie o nim w superlatywach.

5.6.6 Topic Modelling

```
> ldaOut.terms
     Topic 1     Topic 2  Topic 3
[1,] "batteri"   "price"  "use"
[2,] "good"      "camera" "super"
[3,] "recommend" "clear"  "day"
[4,] "great"     "take"   "work"
[5,] "screen"    "even"   "long"
```

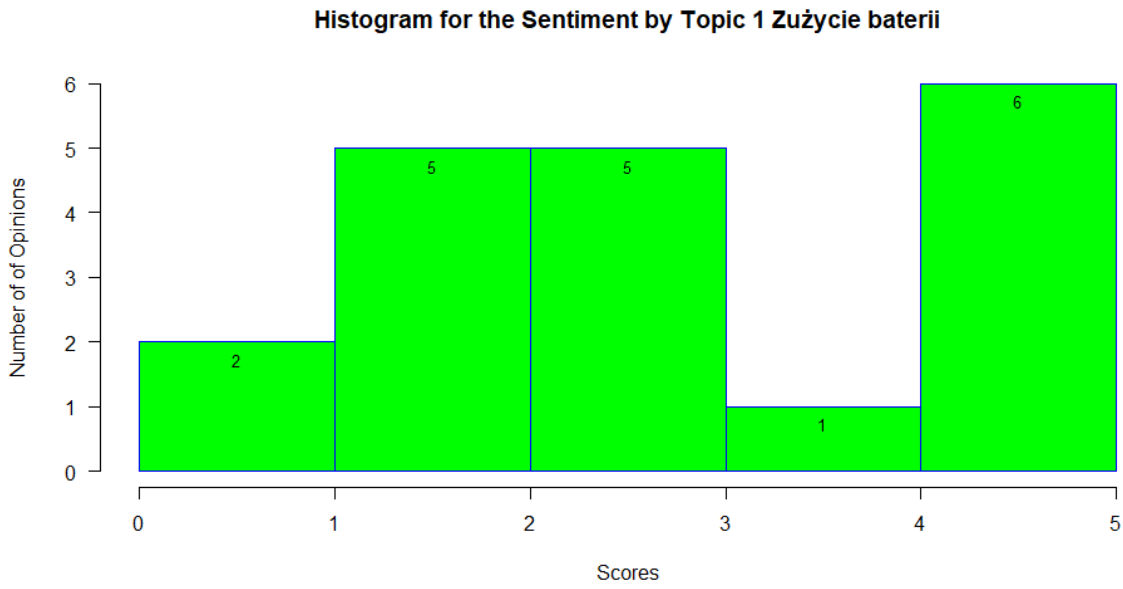
	TOPIC 1	**TOPIC 2**	**TOPIC 3**
NAZWA	Zużycie baterii	Zalety	Wydajność
OPIS	Bardzo dobra pojemność baterii w stosunku do wyświetlacza	Główne plusy telefonu, takie jak cena czy aparat	Temat nawiązuje do długości pracy telefonu oraz jego użytkowania
LICZBA OPINII	19	18	13
NUMER KOMENTARZA	5, 6, 10, 11, 19, 23, 24, 25, 26, 27, 30, 34, 37, 38, 39, 40, 41, 43, 45	3, 4, 9, 13, 15, 16, 17, 18, 21, 22, 28, 32, 35, 44, 47, 48, 49, 50	1, 2, 7, 8, 12, 14, 20, 29, 31, 33, 36, 42, 46

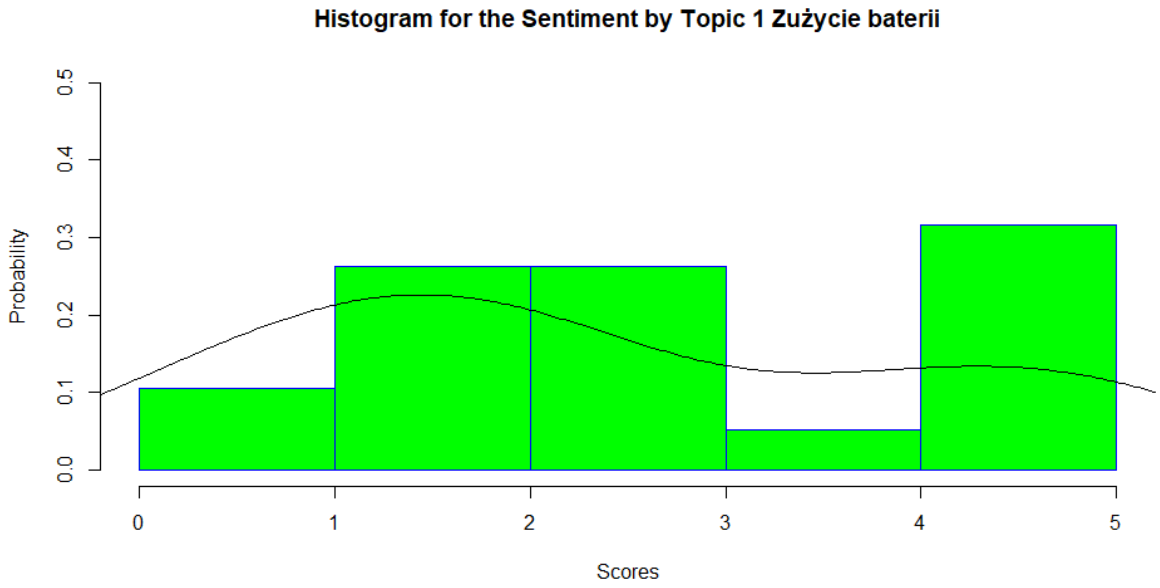
Możemy wyróżnić jeden delikatnie odstający temat pod względem liczby opinii. Dotyczy on wydajności telefonu oraz jego użytkowania i stanowi 26% wszystkich opinii.

TOPIC 1 - Zużycie baterii

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   2.368   4.000   5.000
> minn<-min(m1$Score)
> minn
[1] 0
> maxx<-max(m1$Score)
> maxx
[1] 5
```





```
> pos1$Score
 [1] 2 5 4 4 2 1 1 2 1 2 1 2 1 4 5 5 3
> length(pos1$Score)
[1] 17
```

```
> neu1$Score
[1] 0 0
> length(neu1$Score)
[1] 2
```

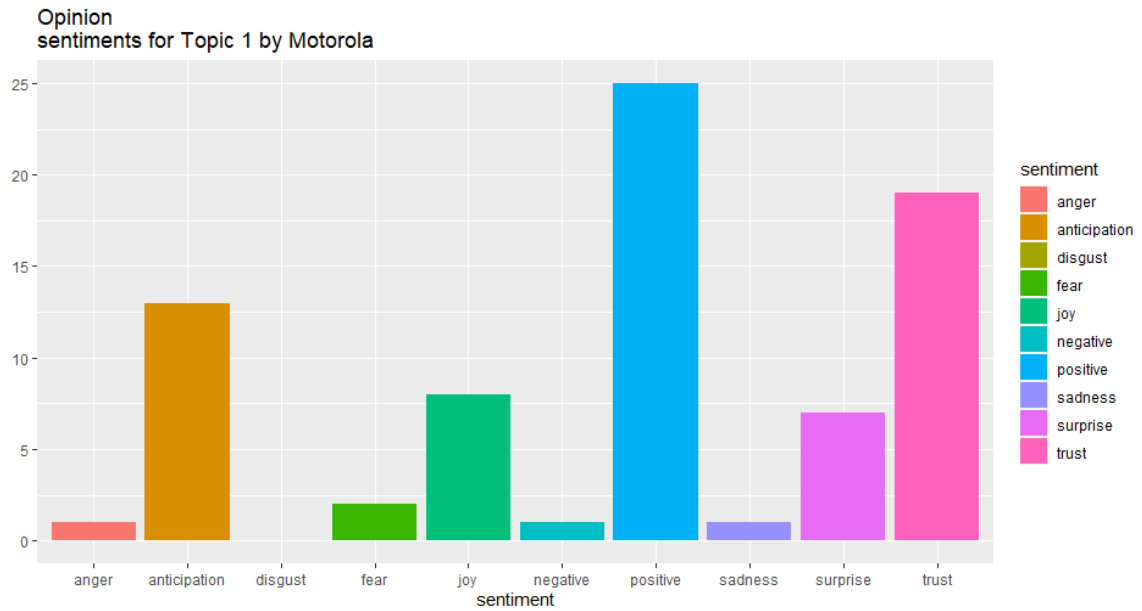
```
> neg1$Score
integer(0)
> length(neg1$Score)
[1] 0
```

Topic 1 dla urządzenia firmy Motorola odnosi się do zużycia baterii. Klienci w tym kontekście zachwalali produkt. Temat ten pojawiał się w największej części opinii skierowanych do tego modelu smartphone'a, a ich pozytywny wydźwięk stanowi fundament do stwierdzenia, że klienci byli bardzo zadowoleni możliwościami jakie oferuję Morotola - Moto G8 Power.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych i neutralnych. Można z nich wnioskować, że klienci najbardziej cenili sobie wydajność baterii, pracy smartphonu, aparatu, ekranu, oraz stosunek ceny do jakości, dzięki czemu chętnie rekomendowali go i pisali o nim w superlatywach. Opinie neutralne również wskazują na pozytywne odczucia, skupiając się ostatecznie na ogólnym zadowoleniu z zakupu i ponownie wskazując baterię jako główny atut telefonu.

- **Opinion sentiments**

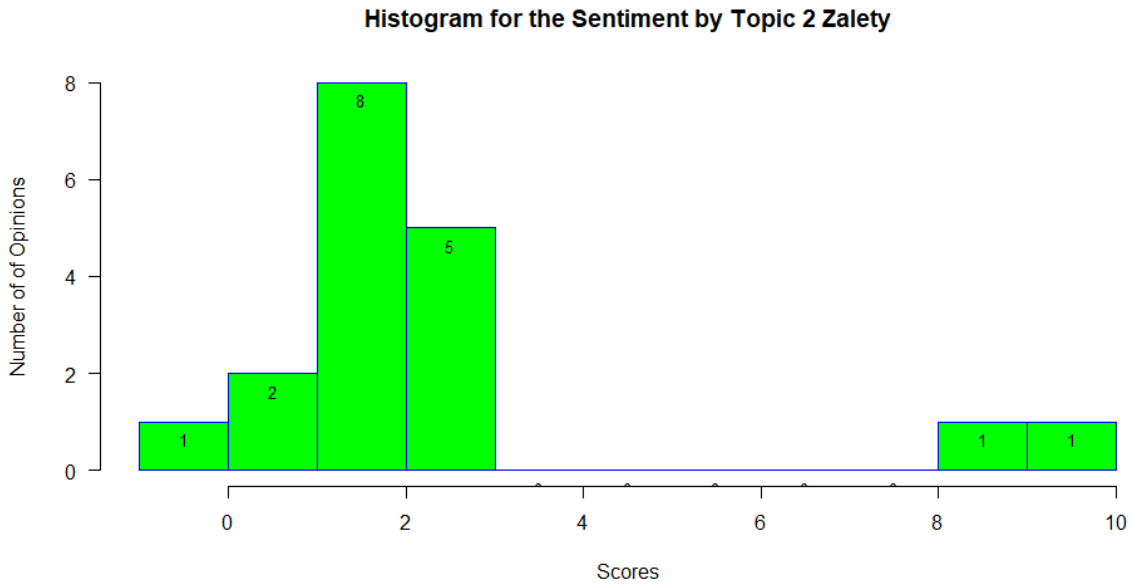


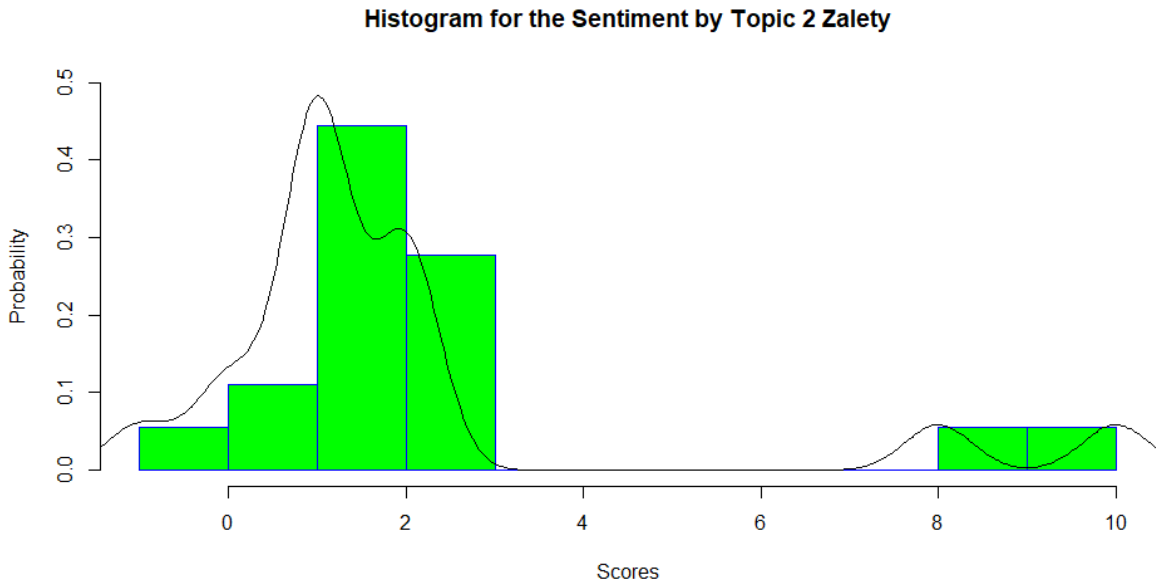
Sentyment dla topic'u 1 oscyluje głównie w odczuciach pozytywnych. Wskaźniki takie jak pozytyw oraz zaufanie do produktu wskazują na główne emocje jakimi kierowali się nabywcy pisząc opinie na temat urządzenia. W tym przypadku Negatywne odczucia charakteryzują się bardzo niskimi bądź zerowymi wartościami, względem pozytywnych.

TOPIC 2 - Zalety

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.000   1.000   1.000   1.944   2.000  10.000
> minn<-min(m1$Score)
> minn
[1] -1
> maxx<-max(m1$Score)
> maxx
[1] 10
```






```
> pos1$Score
 [1]  1  2  1  2 10  1  8  2  2  1  1  1  1  1  2
> length(pos1$Score)
[1] 15
```

```
> neu1$Score
[1] 0 0
> length(neu1$Score)
[1] 2
```

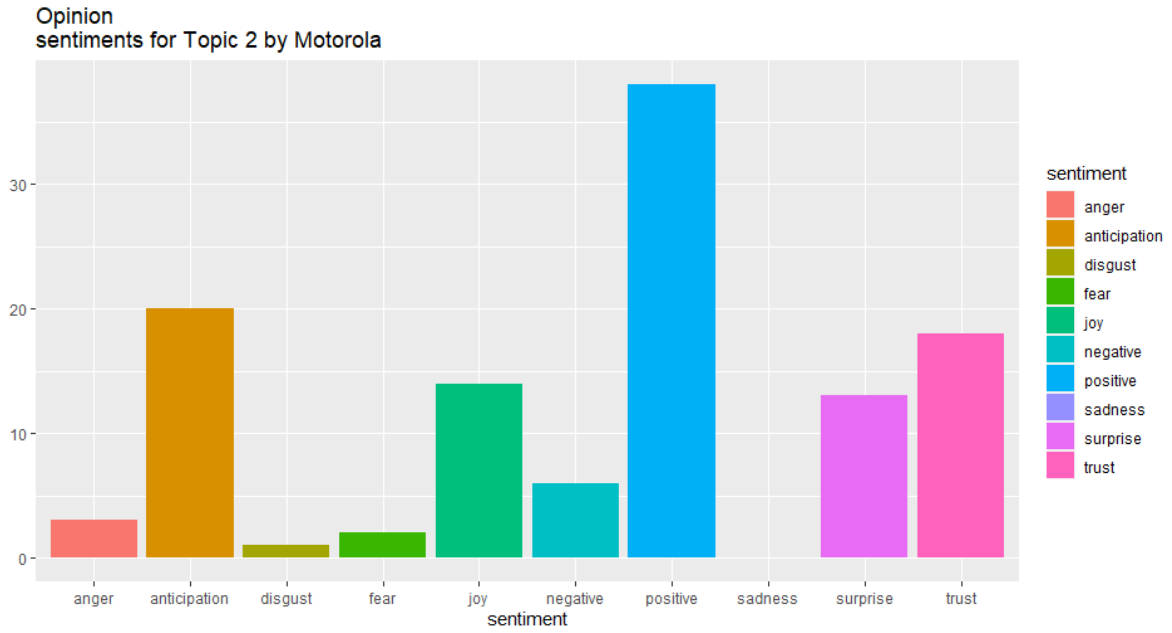
```
> neg1$Score
[1] -1
> length(neg1$Score)
[1] 1
```

Topic 2 skupia się na zaletach jakie wskazywali nabywcy w swoich opiniach. Wydźwięk opinii ma tu charakter głównie pozytywny, osiągnął ich aż 15 na 18, co wskazuję że wiekszość klientów zachwalało produkt. Dwóch klientów było neutralnie nastawionych do produktu, tylko jeden negatywnie.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych, neutralnych i negatywnych. Można z nich wnioskować, że klienci najbardziej cenią szeroki wachlarz funkcji i możliwości telefonu, wydajność baterii, oraz stosunek ceny do jakości. W opiniach neutralnych górowały głosy, iż telefon sprostał oczekiwaniom, zaś w negatywnych - że firma nie odnosi się do stawianych im pytań i zarzutów.

- **Opinion sentiments**

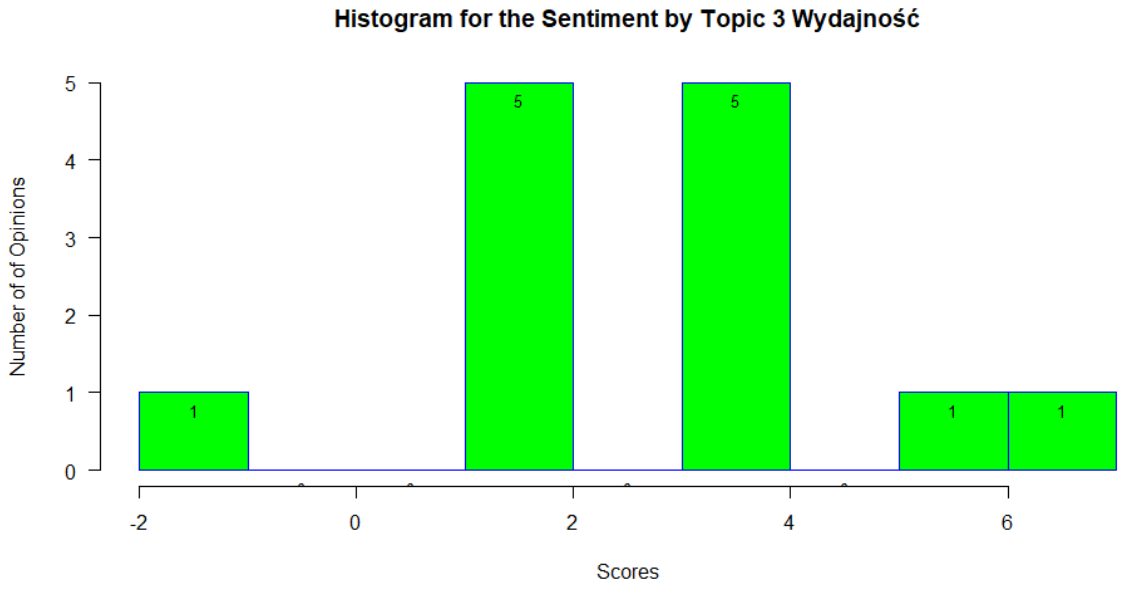


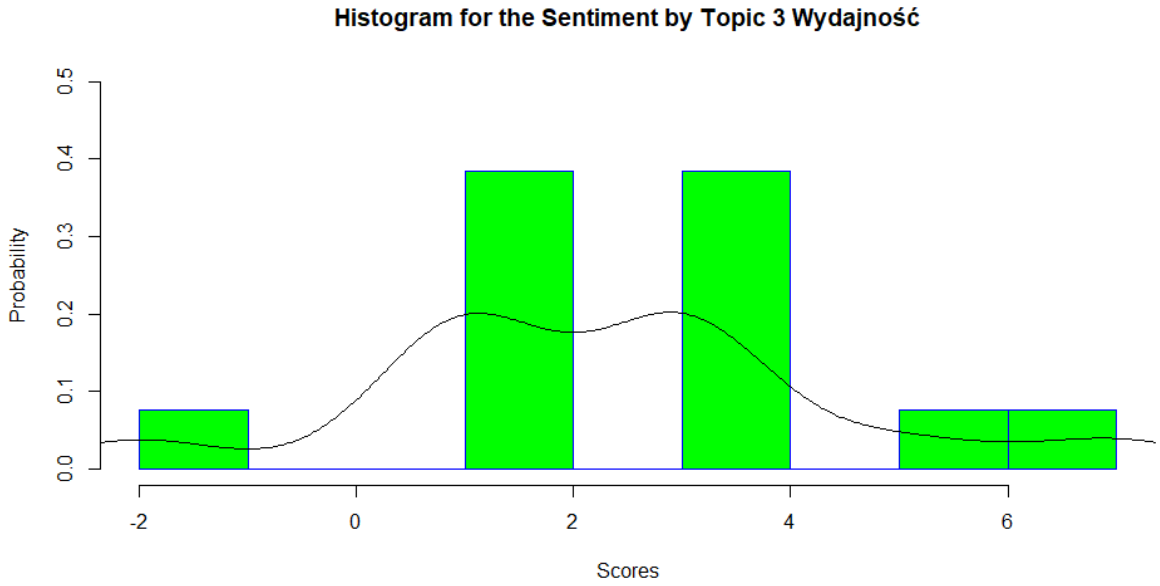
Topic 2 oraz jego sentyment wskazał, że najwięcej opinii zawiera w swojej treści uczucia pozytywne. Największą wartość osiąga słupek odnoszący się do pozytywu, następnie są oczekiwania oraz zaufanie pod względem produktu. Wartości charakteryzujące negatywne odczucia w porównaniu do pozytywnych sa na bardzo niskim poziomie.

TOPIC 3 - Wydajność

- **Scoring**

```
> summary(m1$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -2.000   1.000   3.000   2.308   3.000   7.000
> minn<-min(m1$Score)
> minn
[1] -2
> maxx<-max(m1$Score)
> maxx
[1] 7
```






```
> pos1$Score
 [1] 1 1 7 1 3 5 3 3 1 3 3 1
> length(pos1$Score)
[1] 12
```

```
> neu1$Score
integer(0)
> length(neu1$Score)
[1] 0
```

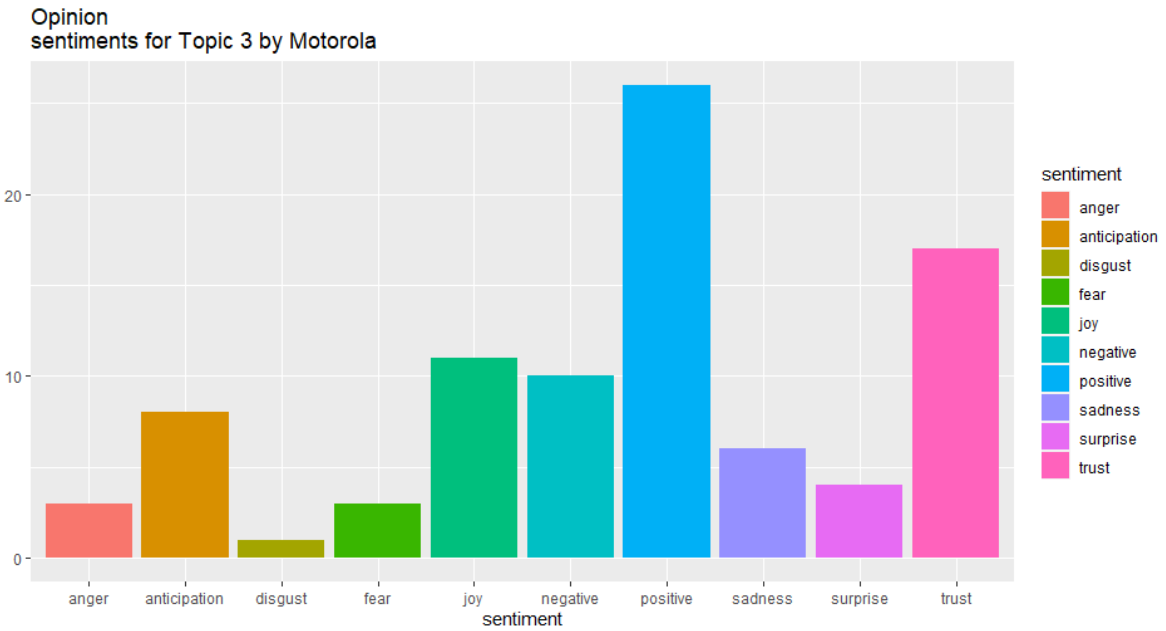
```
> neg1$Score
[1] -2
> length(neg1$Score)
[1] 1
```

Wyniki scoringu dla topicu 3 kształtują się w atmosferze pozytywnej. Zauważa się, że pod względem wydajności model telefonu osiągnął 12 opinii pozytywnych oraz 1 negatywną.

- **Wordcloud**

Powyższe wordcloudy graficznie przedstawiają częstotliwości występowania słów w opiniach pozytywnych i neutralnych. Można z nich wnioskować, że klienci najbardziej cenili sobie wydajność baterii, pracy smartphonu, ekranu, aparatu oraz stosunek ceny do jakości. Opinie neutralne wskazują zaś, iż telefon powinien dać więcej do zaoferowania.

- **Opinion sentiments**



Sentyment dla topic'u 3 wskazał, że najwięcej opinii w tym temacie oscyluje na poziomie emocji pozytywnych. Zauważalne jest to głównie przy dwóch słupkach histogramu: niebieskim – pozytyw oraz z w kolorze magenty – zaufanie. W odróżnieniu jednak od poprzednich przypadków, obserwowalny jest znaczący wzrost odczuć negatywnych.

5.7 Podsumowanie

Analiza korpusu pozwoliła odpowiedzieć na postawione pytania badawcze: w jaki sposób nabywcy wybranych smartphone'ów wypowiadają się na temat swojego zakupu, które elementy smartphonów były najczęściej komentowane przez nabywców oraz który z wybranych modeli smartphonów uzyskał najlepszy wynik względem komentarzy.

Dla każdego telefonu sprawdzono DTM, wyliczono statystyki opisowe, przeprowadzono preprocessing i sprawdzono jak wpłynął on na korpus; utworzono zipf, wordclouds i wykonano topic modeling, którego następstwem była analiza sentymentalna, wordcloud oraz scoring dla każdego topicu z osobna. Tematami, które występowały najczęściej były: użytkowanie/wydajność i kamera.

Na podstawie badania wysunięto wnioski, że klienci najczęściej w swoich opiniach oceniali jakość baterii, robionych zdjęć, jak również aparat i cały wachlarz funkcji w postaci optymistycznego ogółu, wynikiem którego użytkownicy pisali o produktach w superlatywach i polecali ich zakup.

We wszystkich przypadkach zdecydowanie przeważały pozytywne opinie, a najlepiej o produkcie wypowiadali się posiadacze urządzenia Samsung (45), natomiast najwięcej neutralnych oraz negatywnych głosów wystąpiło w grupie użytkowników Xiaomi (w liczbie kolejno: 12 i 4).

Tabela z komentarzami

MARKA	KOMENTARZ			SUMA
	POZYTYWNY	NEUTRALNY	NEGATYWNY	
APPLE	41	7	2	50
XIAOMI	34	12	4	50
HUAWEI	36	9	3	48
SAMSUNG	45	4	1	50
MOTOROLA	44	4	2	50

KODY

```
#Ustalamy ścieżkę
setwd("C:/Users/Karolina/Desktop/Nauka/ZIE
/Data Mining/Opinion_Mining")

#Wgrywamy nasze opinie do R
Data<-read.csv("Opinion_mining.csv",
         header = TRUE,
         sep = ";",
         strip.white = TRUE,
         fill = TRUE,
         comment.char = "#",
         stringsAsFactors = FALSE
)

#Wyświetlamy pierwsze 6 wierszy
head(Data)

#Zamieniamy to na macierz (ramkę danych)
MyData = as.data.frame.matrix(Data)

#Wybieramy dane tylko dla Apple
MyData1<-
MyData[MyData$smartfon=="SAMSUNG SM-
A715 Galaxy",]

#Zostawiamy tylko jedną kolumnę z opiniami
MyData_1 <- MyData1[,4]

#Sprawdzamy liczbę opinii
n<-length(MyData_1)
n

#Znowu sprawdzamy pierwsze 6 wierszy
head(MyData_1)

library(tm)
#Tworzymy korpus
docs <-VCorpus(x = VectorSource(MyData_1),
         readerControl =
list(reader=readPlain,language="en"))
docs

# Tworzymy dtm w celu sprawdzenia
podstawowych informacji (liczba plików czy
ilość słów)
docs <- tm_map(docs, PlainTextDocument)
dtm <- DocumentTermMatrix(docs)
dtm

#Numerujemy nazwy wierszy
rownames(dtm)<-seq(1,n)
rownames(dtm)

#Sprawdzamy jak wygląda 1 opinia
writeLines(as.character(docs [[1]]))

#sumujemy długość komentarzy
doc_length <-
as.data.frame(rowSums(as.matrix(dtm)))

#maksymalna długość
max_length<-max(doc_length)
max_length
#minimalna
min_length<-min(doc_length)
min_length
#średnia
aver_length<-
mean(rowSums(as.matrix(dtm)))
aver_length

#Preprocesing

docs <- tm_map(docs, removePunctuation)   #
usuwa znaki punkcyjne
docs <- tm_map(docs, removeNumbers)   #
usuwa liczby

#Podgląd 1 opinii
writeLines(as.character(docs[[1]]))

#usuwanie znaków, które nam się nie
podobają
for (j in seq(docs)) {
```

```
  docs[[j]] <- gsub("/", " ", docs[[j]])
  docs[[j]] <- gsub("@", " ", docs[[j]])
  docs[[j]] <- gsub("–", " ", docs[[j]])
  docs[[j]] <- gsub("’", " ", docs[[j]])
  docs[[j]] <- gsub("“", " ", docs[[j]])
  docs[[j]] <- gsub("…", " ", docs[[j]])
  docs[[j]] <- gsub("‘", " ", docs[[j]])
  docs[[j]] <- gsub(")", " ", docs[[j]])
  docs[[j]] <- gsub("”", " ", docs[[j]])
}
#znowu podgląd
writeLines(as.character(docs[[1]]))

# zamiana dużych liter na małe
docs <- tm_map(docs, tolower)
writeLines(as.character(docs[[1]]))

# usuwamy stopwords
docs <- tm_map(docs, removeWords,
stopwords("English"))
writeLines(as.character(docs[[1]]))

# wgrywamy plik txt z stopwords
StW<-
read.table("C:/Users/Karolina/Desktop/Nauka
/ZIE/Data Mining/StopWords.txt")
StW

# zamiana wektora na tekst
StWW<-as.character(StW$V1)
StWW

# Usuwamy słowa, które znajdują się w txt
docs <- tm_map(docs, removeWords, StWW)
writeLines(as.character(docs[[1]]))

# Usuwamy spacje, które pojawiły się po
usunięciach
docs <- tm_map(docs, stripWhitespace)
writeLines(as.character(docs[[1]]))

# Stemming
library(SnowballC)
for (j in seq(docs)) {
docs[[j]]<-stemDocument(docs[[j]], language
= "english")
}
writeLines(as.character(docs[[1]]))

#Tworzenie macierzy dtm
docs <- tm_map(docs, PlainTextDocument)

#Mechaniczne usuwanie słów
#długość słowa od 3 do 15 liter, dodatkowo
słowo musi wystąpić w co najmniej 2 opiniach
dtm <-DocumentTermMatrix(docs,
control=list(wordLengths=c(3, 15),bounds =
list(global = c(2,Inf))))
dtm

# Najważniejsze, najczęściej występujące
słowa
freqr <- colSums(as.matrix(dtm))
length(freqr)
freq <- sort(freqr, decreasing=TRUE)   # ilość
słów malejąco
# Pierwsze 15 słów
head(freq, 15)
# Ostatnie 15 słów
tail(freq, 15)

#Prawo Zipfa
freqr <- colSums(as.matrix(dtm))
length(freqr)
mk<-min(head(freq, 12))
mk
wf=data.frame(word=names(freq),freq=freq)

library(ggplot2)
# Zipf's law with minimal frequency = MK
p <- ggplot(subset(wf, freq>mk), aes(x =
reorder(word, -freq), y = freq))
p <- p + geom_bar(stat="identity")
p <- p +
theme(axis.text.x=element_text(angle=45,
hjust=1))
p
```

```
#Wordcloud
library(wordcloud)
dev.new(width = 100, height = 150, unit =
"px")
# Chmura, gdzie min częstotliwość
występowania (frequency) = 10
set.seed(42)
wordcloud(names(freq), freq,
min.freq=8,colors=brewer.pal(6, "Dark2"))

# Chmura, w której mamy 40 najczęściej
występujących słów
set.seed(142)
wordcloud(names(freq), freq, max.words=50,
colors=brewer.pal(6, "Dark2"))

#TOPIC MODELLING
#Sprawdzamy sobie liczbę słów w każdym
wierszu
raw.sum=apply(dtm,1,FUN=sum)
raw.sum

#liczba pustych wierszy -> wytłumaczyć o co
chodzi (np. "ok" znika itp)
mmm<-nrow(dtm[raw.sum==0,])
mmm

# if mmm=0, only create new matrix dtm2 and
NN (number of rows in DTM)
# if mmm>0, delete the rows with zero's
sum from dtm
if (mmm==0) {
 dtm2<-dtm
 NN<-nrow(dtm)
 NN
} else {
 dtm2<-dtm[raw.sum!=0,]
 NN<-nrow(dtm2)
}

#liczba opinii przed usunięciem pustych
wierszy
n
# i po usunięciu
NN

#nowa macierz dtm2
dtm2

#LDA Topic Modelling
# Jeśli best=TRUE zwracany jest tylko
najlepszy model ze wszystkich przebiegów w
# odniesieniu do logarytmu
prawdopodobieństwa

library(topicmodels)
burnin <- 4000
iter <- 2000
thin <- 500
seed <-list(2003,5,63,100001,765)
nstart <- 5
best <- TRUE

# Liczba możliwych tematów
k <- 5
#teraz tworzymy te tematy
ldaOut <-LDA(dtm2, k, method="Gibbs",
control=list(nstart=nstart, seed = seed,
best=best,
burnin = burnin,
iter = iter,
thin=thin))
str(ldaOut)
ldaOut.terms <- as.matrix(terms(ldaOut,5))
ldaOut.terms

# Tworzymy macierz, w której znajdują się
prawdopodobieństwa związane z każdym
przypisaniem tematu
topicProbabilities <-
as.data.frame(ldaOut@gamma)
topicProbabilities

#Obliczamy "proporcje" (podobieństwo ) dla
każdego topiku względem korpusu
col.sum=apply(topicProbabilities,2,FUN=sum)
# 1 - rows sum, 2- columns sum
col.sum<-as.matrix(col.sum)
dim(col.sum)
sum.TP=col.sum/sum(col.sum)
sum.TP
```

```
#Przypisujemy temat do każdego wiersza
ldaOut.topics <- as.matrix(topics(ldaOut))
number_comments <- length(ldaOut.topics)
number_comments
ss <- seq(1,number_comments, by=1)
comment_n <- paste("Comment",
as.character(ss), sep = " ", collapse = NULL)
length(comment_n)
rownames(ldaOut.topics) <- comment_n

#wyświetliłam pierwsze 6 komentarzy tak dla
podglądu
head(ldaOut.topics)
#write.csv(ldaOut.topics,file=paste("LDAGibbs
",k, "DocsToTopics.csv"))

#CREATING TOPIC ORIENTED CORPA

#Topics Data Frame building
nrow(ldaOut.topics)
Comment<-seq(1, NN, by=1)
Comment
wf=data.frame(Comment=Comment,
Topics=ldaOut.topics)
wf

#Uwaga --> tą część powtarzamy dla każdego
tematu

#Building Sub-Corpus of Topic 1
#Wybieramy tylko te wiersze, w których w 2
kolumnie jest 1 (tzw Topic 1)
topic1<-wf[wf[2] == 1,]
topic1$Comment
#liczba opini z topiku 1
kk1<-nrow(topic1)
kk1
#ogólna liczba opini
kk<-nrow(dtm2)
kk

#tworzymy listę (wektor) z numerkami opini z
tematu 1
list1<-c()
i=1
while(i<=kk) {
 if (wf[i,2]==1) { # Find Topic 1
        list1<-c(list1,i)}
 i=i+1
}
list1
# tworzymy ramkę danych z opiniami
przypisanymi do tematu 1
wf1=NULL
for (i in 1:kk) {
 for (j in 1:kk1) {
        if (i==list1[j]){
        c <-
data.frame(file=as.character(wf[list1[j],1]),doc
ument=as.character(docs[[i]]))
        wf1=rbind(wf1,c)
        }
 }
}
wf1
wf1$document[1]

#Corpus Creating
# Corpus for Topic 1
Topic_1_docs <-
Corpus(VectorSource(as.character(wf1$docu
ment)))
writeLines(as.character(Topic_1_docs [[1]]))
Topic_1_docs

#Writing in CSV File_ with Comments of Topic
1
mycorpus_dataframe <-
data.frame(text=wf1$document,
stringsAsFactors=F)
mycorpus_dataframe
write.csv(mycorpus_dataframe,
"Topic_1_docs.csv", row.names=FALSE)

#Tworzenie dtm, prawa Zipfa i chmurki dla
topiku
dtm_topic_1 <-
DocumentTermMatrix(Topic_1_docs)
dtm_topic_1
```

```
#SENTIMENT ANALYSIS
install.packages("syuzhet",
dependencies=TRUE,
repos='http://cran.rstudio.com/')
library("plyr")
library("stringr")
library("syuzhet")
neg=scan("negative-words.txt",
what="character", comment.char=";" )
pos=scan("positive-words.txt",
what="character", comment.char=";" )

#Initialization of the Sentiment analysis
Procedure
#szukamy czy słowa w naszym dokumencie są
pozytywne czy negatywne
score.sentiment = function(docs, pos.words,
neg.words, .progress='none')
{
  scores = laply(docs_s, function(docs,
pos.words, neg.words) {
          word.list = str_split(docs, '\\s+')
          words = unlist(word.list)
# compare our words to the dictionaries of
positive & negative terms
          pos.matches = match(words,
pos.words)
          neg.matches = match(words,
neg.words)
# match() returns the position of the matched
term or NA
# we just want a TRUE/FALSE:
          pos.matches = !is.na(pos.matches)
          neg.matches = !is.na(neg.matches)
# and conveniently enough, TRUE/FALSE will
be treated as 1/0 by sum():
          score = sum(pos.matches) -
sum(neg.matches)
          return(score)
  }, pos.words, neg.words, .progress=.progress
)
 scores.df = data.frame(score=scores,
text=docs)
  return(scores.df)
}

#Koniec tej części tu znowu trzeba puszczać
dla różnych tematów

#Topic 1 Sentiment Scoring
result=c()
docs<- Topic_1_docs
m1=c()
for (j in seq(docs)) {
  docs_s=as.character(docs[[j]])
  print(docs_s)
  result = score.sentiment(docs_s, pos, neg)
  newRow1 <- data.frame(Doc=j,Score =
result$score, Documents = result$text)
  #print(newRow1)
  m1<- rbind(m1,newRow1)
  #print(m1)
}
m1[1:3,]

#Statystyka dotycząca scoringu
summary(m1$Score)
minn<-min(m1$Score)
minn
maxx<-max(m1$Score)
maxx

#Histogram
mmm<-maxx-minn
mmm

h<-hist(m1$Score,
          main="Histogram for the Sentiment
by Topic 1 Zdjęcia",
          xlab="Scores",
          ylab="Number of of Opinions",
          right=FALSE,
          border="blue",
      col="green",
          freq=TRUE,
          las=1,
          xlim=c(minn,maxx),
          breaks=mmm
)
```

```
#opisy na histogramie
text(h$mids,h$counts,labels=h$counts,
adj=c(0.5, -0.5),cex = 0.8, pos = 1)
#Informacje z histogramu
m1$Score
h$count

#Histogram nr 2
hist(m1$Score,
 main="Histogram for the Sentiment by Topic
1 Zdjecia",
          xlab="Scores",
          ylab="Probability",
          border="blue",
          col="green",
          prob = TRUE,
          right=FALSE,
          xlim=c(minn,maxx),
          ylim=c(0,0.5),
          breaks=mmm
)
lines(density(m1$Score))

m11<-as.matrix(m1)
m11

#Dzielimy temat 1 na pozytywne, negatywne i
neutralne opinie

#Positive - score >=1
pos1<-m1[m1$Score>=1,]
pos1$Documents
pos1$Score
length(pos1$Score)

#Neutral - score <1 and >=0
neu1<-m1[(m1$Score<1)&(m1$Score>=0),]
neu1$Documents
neu1$Score
length(neu1$Score)

#Negative - score <0
neg1<-m1[m1$Score<0,]
neg1$Score
length(neg1$Score)
neg1$Documents

#Robimy z tego podkorpusy
pos_docs_1 <-
Corpus(VectorSource(pos1$Documents))
pos_docs_1
neu_docs_1 <-
Corpus(VectorSource(neu1$Documents))
neu_docs_1
neg_docs_1 <-
Corpus(VectorSource(neg1$Documents))
neg_docs_1

#Zamieniamy to na macierze
#Writing in CSV File_ with Positive Comments
pos_docs_1_dataframe <-
data.frame(text=pos1$Documents,
stringsAsFactors=F)
pos_docs_1_dataframe

#Writing in CSV File_ with Negative
Comments
neg_docs_1_dataframe <-
data.frame(text=neg1$Documents,
stringsAsFactors=F)
neg_docs_1_dataframe

#Writing in CSV File_ with Neutral Comments
neu_docs_1_dataframe <-
data.frame(text=neu1$Documents,
stringsAsFactors=F)
neu_docs_1_dataframe

#Zipf i chmura dla pozytywnych - trzeba
jeszcze zrobić dla neutralnych itp
dtm_topic_1_pos <-
DocumentTermMatrix(neg_docs_1)
dtm_topic_1_pos
freqr <- colSums(as.matrix(dtm_topic_1_pos))
length(freqr)
freq <- sort(freqr, decreasing=TRUE)
mk<-min(head(freq, 30))
mk
wf=data.frame(word=names(freq),freq=freq)
```

```
# wordcloud
dev.new(width = 100, height = 100, unit =
"px")
set.seed(142)
dark2 <- brewer.pal(6, "Dark2")
wordcloud(names(freq), freq,
max.words=100, rot.per=0.2, colors=dark2)

#SENTIMENT ANALYSIS USING NRC WORD-
EMOTION ASSOCIATION LEXICON

#wczytujemy plik stworzony dla 1 Topicu
Topic_1<-read.csv("Topic_1_docs.csv",
        header = TRUE,
        sep = ",", # or ";"
        strip.white = TRUE,
        fill = TRUE,
        comment.char = "#",
        stringsAsFactors = FALSE
)
head(Topic_1)

#Tworzymy ramkę danych
Topic_1 = as.data.frame.matrix(Topic_1)
mycorpus_dataframe1<-
data.frame(text=Topic_1, stringsAsFactors=F)
mycorpus_dataframe1

#Liczy Sentyment z całego tekstu przez inny
słownik - szczegóły emocje

#proporcje ile czego było
usableText=str_replace_all(mycorpus_datafra
me1$text,"[^[:graph:]]", " ")
d<-get_nrc_sentiment(usableText)
head(d)
d$anger

t(d)
td<-data.frame(t(d))
td[,5]
td_new <- data.frame(rowSums(td))
td_new
names(td_new)[1] <- "count"
td_new <- cbind("sentiment" =
rownames(td_new), td_new)
rownames(td_new) <- NULL
td_new
td_new2<-td_new[1:10,]
td_new2

#Wykres
qplot(sentiment, data=td_new2,
weight=count,
geom="bar",fill=sentiment)+ggtitle("Opinion
sentiments for Topic 1 by Apple")
```